



Search  for

Limits Preview/Index History Clipboard Details

Display  Show  Sort by  Send to

About Entrez

Text Version

All: 100 Review: 0

Items 1 - 20 of 100

Page  of 5 Next

#### Entrez PubMed

Overview

Help | FAQ

Tutorials

New/Noteworthy

E-Utilities

#### PubMed Services

Journals Database

MeSH Database

Single Citation Matcher

Batch Citation Matcher

Clinical Queries

Special Queries

LinkOut

My NCBI

#### Related Resources

Order Documents

NLM Mobile

NLM Catalog

NLM Gateway

TOXNET

Consumer Health

Clinical Alerts

ClinicalTrials.gov

PubMed Central

☐ **1:** Hutchinson GB, Hayden MR. [Related Articles, Links](#)



The prediction of exons through an analysis of spliceable open reading frames.

Nucleic Acids Res. 1992 Jul 11;20(13):3453-62.

PMID: 1321415 [PubMed - indexed for MEDLINE]

☐ **2:** Solovyev VV, Salamov AA, Lawrence CB. [Related Articles, Links](#)



Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames.

Nucleic Acids Res. 1994 Dec 11;22(24):5156-63.

PMID: 7816600 [PubMed - indexed for MEDLINE]

☐ **3:** Solovyev VV, Salamov AA, Lawrence CB. [Related Articles, Links](#)



The prediction of human exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames.

Proc Int Conf Intell Syst Mol Biol. 1994;2:354-62.

PMID: 7584412 [PubMed - indexed for MEDLINE]

☐ **4:** Zhang DL, Ji L, Li YD. [Related Articles, Links](#)



[Analysis, identification and correction of some errors of model refseqs appeared in NCBI Human Gene Database by in silico cloning and experimental verification of novel human genes]

Yi Chuan Xue Bao. 2004 May;31(5):431-43. Chinese.

PMID: 15478601 [PubMed - indexed for MEDLINE]

☐ **5:** Xu Y, Mural R, Shah M, Uberbacher E. [Related Articles, Links](#)



Recognizing exons in genomic sequence using GRAIL II.

Genet Eng (N Y). 1994;16:241-53.

PMID: 7765200 [PubMed - indexed for MEDLINE]

☐ **6:** Chen T, Zhang MQ. [Related Articles, Links](#)



Pombe: a gene-finding and exon-intron structure prediction system for fission yeast.

Yeast. 1998 Jun 15;14(8):701-10.

PMID: 9675815 [PubMed - indexed for MEDLINE]

☐ **7:** Thanaraj TA. [Related Articles, Links](#)



Positional characterisation of false positives from computational prediction

of human splice sites.

Nucleic Acids Res. 2000 Feb 1;28(3):744-54.

PMID: 10637326 [PubMed - indexed for MEDLINE]

- ☐ **8:** [Solovyev VV, Salamov AA, Lawrence CB.](#)

[Related Articles, Links](#)



Identification of human gene structure using linear discriminant functions and dynamic programming.

Proc Int Conf Intell Syst Mol Biol. 1995;3:367-75.

PMID: 7584460 [PubMed - indexed for MEDLINE]

- ☐ **9:** [Gottlieb LD, Ford VS.](#)

[Related Articles, Links](#)



The 5' leader of plant PgiC has an intron: the leader shows both the loss and maintenance of constraints compared with introns and exons in the coding region.

Mol Biol Evol. 2002 Sep;19(9):1613-23.

PMID: 12200488 [PubMed - indexed for MEDLINE]

- ☐ **10:** [Farber R, Lapedes A, Sirotkin K.](#)

[Related Articles, Links](#)



Determination of eukaryotic protein coding regions using neural networks and information theory.

J Mol Biol. 1992 Jul 20;226(2):471-9.

PMID: 1640461 [PubMed - indexed for MEDLINE]

- ☐ **11:** [Yu W, Ikeda M, Abe H, Honma S, Ebisawa T, Yamauchi T, Honma K, Nomura M.](#)

[Related Articles, Links](#)



Characterization of three splice variants and genomic organization of the mouse BMAL1 gene.

Biochem Biophys Res Commun. 1999 Jul 14;260(3):760-7.

PMID: 10403839 [PubMed - indexed for MEDLINE]

- ☐ **12:** [Liu HX, Cartegni L, Zhang MQ, Krainer AR.](#)

[Related Articles, Links](#)



A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes.

Nat Genet. 2001 Jan;27(1):55-8.

PMID: 11137998 [PubMed - indexed for MEDLINE]

- ☐ **13:** [Chyan YJ, Strauss PR, Wood TG, Wilson SH.](#)

[Related Articles, Links](#)



Identification of novel mRNA isoforms for human DNA polymerase beta.

DNA Cell Biol. 1996 Aug;15(8):653-9.

PMID: 8769567 [PubMed - indexed for MEDLINE]

- ☐ **14:** [Meyer IM, Durbin R.](#)

[Related Articles, Links](#)



Gene structure conservation aids similarity based gene prediction.

Nucleic Acids Res. 2004 Feb 4;32(2):776-83. Print 2004.

PMID: 14764925 [PubMed - indexed for MEDLINE]

- ☐ **15:** [Besemer J, Lomsadze A, Borodovsky M.](#)

[Related Articles, Links](#)



GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions.

Nucleic Acids Res. 2001 Jun 15;29(12):2607-18.

PMID: 11410670 [PubMed - indexed for MEDLINE]

- ☐ **16:** [Fukunishi Y, Hayashizaki Y.](#)

[Related Articles, Links](#)



Amino acid translation program for full-length cDNA sequences with frameshift errors.

Physiol Genomics. 2001 Mar 8;5(2):81-7.

PMID: 11242592 [PubMed - indexed for MEDLINE]

☐ **17:** [Roy K, Mitsugi K, Sirotnak FM.](#)

[Related Articles, Links](#)



Additional organizational features of the murine folylpolyglutamate synthetase gene. Two remotely situated exons encoding an alternate 5' end and proximal open reading frame under the control of a second promoter.

J Biol Chem. 1997 Feb 28;272(9):5587-93.

PMID: 9038166 [PubMed - indexed for MEDLINE]

☐ **18:** [Solovyev VV, Lawrence CB.](#)

[Related Articles, Links](#)



Identification of human gene functional regions based on oligonucleotide composition.

Proc Int Conf Intell Syst Mol Biol. 1993;1:371-9.

PMID: 7584359 [PubMed - indexed for MEDLINE]

☐ **19:** [Miriami E, Motro U, Sperling J, Sperling R.](#)

[Related Articles, Links](#)



Conservation of an open-reading frame as an element affecting 5' splice site selection.

J Struct Biol. 2002 Oct-Dec;140(1-3):116-22.

PMID: 12490159 [PubMed - indexed for MEDLINE]

☐ **20:** [Rossi AM, Tate AD, van Zeeland AA, Vrieling H.](#)

[Related Articles, Links](#)



Molecular analysis of mutations affecting hprt mRNA splicing in human T-lymphocytes in vivo.

Environ Mol Mutagen. 1992;19(1):7-13.

PMID: 1732105 [PubMed - indexed for MEDLINE]

Items 1 - 20 of 100

Page

1

of 5 Next

Display

Summary



Show

20



Sort by



Send to



[Write to the Help Desk](#)

[NCBI](#) | [NLM](#) | [NIH](#)

[Department of Health & Human Services](#)

[Privacy Statement](#) | [Freedom of Information Act](#) | [Disclaimer](#)

Aug 14 2006 08:07:58



A service of the National Library of Medicine  
and the National Institutes of Health

My NCBI  
[\[Sign In\]](#) [\[Regis\]](#)

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Bool

Search PubMed for

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Sort by Send to

About Entrez

Text Version

All: 100 Review: 0

Items 21 - 40 of 100

Previous  2 of 5 Next

Entrez PubMed

Overview

Help | FAQ

Tutorials

New/Noteworthy

E-Utilities

PubMed Services

Journals Database

MeSH Database

Single Citation Matcher

Batch Citation Matcher

Clinical Queries

Special Queries

LinkOut

My NCBI

Related Resources

Order Documents

NLM Mobile

NLM Catalog

NLM Gateway

TOXNET

Consumer Health

Clinical Alerts

ClinicalTrials.gov

PubMed Central

☐ 21: [Eden E, Brunak S.](#)

[Related Articles, Links](#)



Analysis and recognition of 5' UTR intron splice sites in human pre-mRNA.

Nucleic Acids Res. 2004 Feb 11;32(3):1131-42. Print 2004.

PMID: 14960723 [PubMed - indexed for MEDLINE]

☐ 22: [Liang H, Landweber LF.](#)

[Related Articles, Links](#)



A genome-wide study of dual coding regions in human alternatively spliced genes.

Genome Res. 2006 Feb;16(2):190-6. Epub 2005 Dec 19.

PMID: 16365380 [PubMed - indexed for MEDLINE]

☐ 23: [Min XJ, Butler G, Storms R, Tsang A.](#)

[Related Articles, Links](#)



OrfPredictor: predicting protein-coding regions in EST-derived sequences.

Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W677-80.

PMID: 15980561 [PubMed - indexed for MEDLINE]

☐ 24: [Gao F, Zhang CT.](#)

[Related Articles, Links](#)



Comparison of various algorithms for recognizing short coding sequences of human genes.

Bioinformatics. 2004 Mar 22;20(5):673-81. Epub 2004 Feb 5.

PMID: 14764563 [PubMed - indexed for MEDLINE]

☐ 25: [Kleffe J, Hermann K, Vahrson W, Wittig B, Brendel V.](#)

[Related Articles, Links](#)



GeneGenerator--a flexible algorithm for gene prediction and its application to maize sequences.

Bioinformatics. 1998;14(3):232-43.

PMID: 9614266 [PubMed - indexed for MEDLINE]

☐ 26: [Dietz HC, Kendzior RJ Jr.](#)

[Related Articles, Links](#)



Maintenance of an open reading frame as an additional level of scrutiny during splice site selection.

Nat Genet. 1994 Oct;8(2):183-8.

PMID: 7842017 [PubMed - indexed for MEDLINE]

☐ 27: [Snyder EE, Stormo GD.](#)

[Related Articles, Links](#)



Identification of coding regions in genomic DNA sequences: an

application of dynamic programming and neural networks.

Nucleic Acids Res. 1993 Feb 11;21(3):607-13.

PMID: 8441672 [PubMed - indexed for MEDLINE]

☐ **28:** [Hatzinikolas G, Gibson MA.](#)

[Related Articles, Links](#)



The exon structure of the human MAGP-2 gene. Similarity with the MAGP-1 gene is confined to two exons encoding a cysteine-rich region.

J Biol Chem. 1998 Nov 6;273(45):29309-14.

PMID: 9792630 [PubMed - indexed for MEDLINE]

☐ **29:** [Qiu YH, Chen CN, Malone T, Richter L, Beckendorf SK, Davis RL.](#)

[Related Articles, Links](#)



Characterization of the memory gene dunce of *Drosophila melanogaster*.

J Mol Biol. 1991 Dec 5;222(3):553-65.

PMID: 1660926 [PubMed - indexed for MEDLINE]

☐ **30:** [Shields DC, Higgins DG, Sharp PM.](#)

[Related Articles, Links](#)



GCWIND: a microcomputer program for identifying open reading frames according to codon positional G+C content.

Comput Appl Biosci. 1992 Oct;8(5):521-3.

PMID: 1422886 [PubMed - indexed for MEDLINE]

☐ **31:** [Issac B, Raghava GP.](#)

[Related Articles, Links](#)



EGPred: prediction of eukaryotic genes using ab initio methods after combining with sequence similarity approaches.

Genome Res. 2004 Sep;14(9):1756-66.

PMID: 15342559 [PubMed - indexed for MEDLINE]

☐ **32:** [Gelfand MS.](#)

[Related Articles, Links](#)



Statistical analysis and prediction of the exonic structure of human genes.

J Mol Evol. 1992 Sep;35(3):239-52.

PMID: 1518091 [PubMed - indexed for MEDLINE]

☐ **33:** [Shiokawa D, Tanuma S.](#)

[Related Articles, Links](#)



Cloning of cDNAs encoding porcine and human DNase II.

Biochem Biophys Res Commun. 1998 Jun 29;247(3):864-9.

PMID: 9647784 [PubMed - indexed for MEDLINE]

☐ **34:** [Cabanillas AM, Darling DS.](#)

[Related Articles, Links](#)



Alternative splicing gives rise to two isoforms of Zfphep, a zinc finger/homeodomain protein that binds T3-response elements.

DNA Cell Biol. 1996 Aug;15(8):643-51.

PMID: 8769566 [PubMed - indexed for MEDLINE]

☐ **35:** [Sulekova Z, Reina-Sanchez J, Ballhausen WG.](#)

[Related Articles, Links](#)



Multiple APC messenger RNA isoforms encoding exon 15 short open reading frames are expressed in the context of a novel exon 10A-derived sequence.

Int J Cancer. 1995 Nov 3;63(3):435-41.

PMID: 7591245 [PubMed - indexed for MEDLINE]

☐ **36:** [Poulin F, Brueschke A, Sonenberg N.](#)

[Related Articles, Links](#)



Gene fusion and overlapping reading frames in the mammalian genes for 4E-BP3 and MASK.

J Biol Chem. 2003 Dec 26;278(52):52290-7. Epub 2003 Oct 13.

PMID: 14557257 [PubMed - indexed for MEDLINE]

☐ **37:** [Guigo R.](#)

[Related Articles](#), [Links](#)



Assembling genes from predicted exons in linear time with dynamic programming.

J Comput Biol. 1998 Winter;5(4):681-702.

PMID: 10072084 [PubMed - indexed for MEDLINE]

☐ **38:** [Renshaw RW, Casey JW.](#)

[Related Articles](#), [Links](#)



Transcriptional mapping of the 3' end of the bovine syncytial virus genome.

J Virol. 1994 Feb;68(2):1021-8.

PMID: 8289332 [PubMed - indexed for MEDLINE]

☐ **39:** [Thomson SA, Wallace MR.](#)

[Related Articles](#), [Links](#)



RT-PCR splicing analysis of the NF1 open reading frame.

Hum Genet. 2002 May;110(5):495-502. Epub 2002 Apr 4.

PMID: 12073021 [PubMed - indexed for MEDLINE]

☐ **40:** [Shmatkov AM, Melikyan AA, Chernousko FL, Borodovsky M.](#)

[Related Articles](#), [Links](#)



Finding prokaryotic genes by the 'frame-by-frame' algorithm: targeting gene starts and overlapping genes.

Bioinformatics. 1999 Nov;15(11):874-86.

PMID: 10743554 [PubMed - indexed for MEDLINE]

Items 21 - 40 of 100

Previous **Page**  of 5 Next

Display  Show  Sort by  Send to

[Write to the Help Desk](#)

[NCBI](#) | [NLM](#) | [NIH](#)

[Department of Health & Human Services](#)

[Privacy Statement](#) | [Freedom of Information Act](#) | [Disclaimer](#)

Aug 14 2006 08:07:58



A service of the National Library of Medicine  
and the National Institutes of Health

My NCBI  
[Sign In] [Regis]

All Databases

PubMed

Nucleotide

Protein

Genome

Structure

OMIM

PMC

Journals

Book

Search  for

Limits

Preview/Index

History

Clipboard

Details

Display  Show  Sort by  Send to

About Entrez

Text Version

All: 100 Review: 0

Items 41 - 60 of 100

Previous  3 of 5 Next

## Entrez PubMed

Overview

Help | FAQ

Tutorials

New/Noteworthy

E-Utilities

## PubMed Services

Journals Database

MeSH Database

Single Citation Matcher

Batch Citation Matcher

Clinical Queries

Special Queries

LinkOut

My NCBI

## Related Resources

Order Documents

NLM Mobile

NLM Catalog

NLM Gateway

TOXNET

Consumer Health

Clinical Alerts

ClinicalTrials.gov

PubMed Central

☐ 41: Asker CE, Magnusson KP, Piccoli SP, Andersson K, Klein G, Cole MD, Wiman KG. Related Articles, Links



Mouse and rat B-myc share amino acid sequence homology with the c-myc transcriptional activator domain and contain a B-myc specific carboxy terminal region.

Oncogene. 1995 Nov 16;11(10):1963-9.

PMID: 7478514 [PubMed - indexed for MEDLINE]

☐ 42: Gelfand MS. Related Articles, Links



Computer prediction of the exon-intron structure of mammalian pre-mRNAs.

Nucleic Acids Res. 1990 Oct 11;18(19):5865-9.

PMID: 2216778 [PubMed - indexed for MEDLINE]

☐ 43: Rosby O, Alestrom P, Berg K. Related Articles, Links



Sequence conservation in kringle IV-type 2 repeats of the LPA gene.

Atherosclerosis. 2000 Feb;148(2):353-64.

PMID: 10657572 [PubMed - indexed for MEDLINE]

☐ 44: Schuchman EH, Suchi M, Takahashi T, Sandhoff K, Desnick RJ. Related Articles, Links



Human acid sphingomyelinase. Isolation, nucleotide sequence and expression of the full-length and alternatively spliced cDNAs.

J Biol Chem. 1991 May 5;266(13):8531-9.

PMID: 1840600 [PubMed - indexed for MEDLINE]

☐ 45: Salamov AA, Solovyev VV. Related Articles, Links



Ab initio gene finding in Drosophila genomic DNA.

Genome Res. 2000 Apr;10(4):516-22.

PMID: 10779491 [PubMed - indexed for MEDLINE]

☐ 46: Zhang DL, Li YD, Ji L. Related Articles, Links











[Correction of five different types of errors of model REFSEQs appeared in NCBI human gene database only by using two novel human genes C17orf32 and ZNF362]

Yi Chuan Xue Bao. 2004 Apr;31(4):325-34. Chinese.

PMID: 15487498 [PubMed - indexed for MEDLINE]

☐ 47: Tomita M, Shimizu N, Brutlag DL. Related Articles, Links

Introns and reading frames: correlation between splicing sites and their

-  codon positions.  
Mol Biol Evol. 1996 Nov;13(9):1219-23.  
PMID: 8896374 [PubMed - indexed for MEDLINE]
- ☐ 48: [Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouze P, Brunak S.](#) Related Articles, Links  
Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information.  
Nucleic Acids Res. 1996 Sep 1;24(17):3439-52.  
PMID: 8811101 [PubMed - indexed for MEDLINE]
- ☐ 49: [Brunak S, Engelbrecht J, Knudsen S.](#) Related Articles, Links  
 Prediction of human mRNA donor and acceptor sites from the DNA sequence.  
J Mol Biol. 1991 Jul 5;220(1):49-65.  
PMID: 2067018 [PubMed - indexed for MEDLINE]
- ☐ 50: [Xu Y, Mural RJ, Uberbacher EC.](#) Related Articles, Links  
 Constructing gene models from accurately predicted exons: an application of dynamic programming.  
Comput Appl Biosci. 1994 Dec;10(6):613-23.  
PMID: 7704660 [PubMed - indexed for MEDLINE]
- ☐ 51: [Monaco L, Vicini E, Conti M.](#) Related Articles, Links  
 Structure of two rat genes coding for closely related rolipram-sensitive cAMP phosphodiesterases. Multiple mRNA variants originate from alternative splicing and multiple start sites.  
J Biol Chem. 1994 Jan 7;269(1):347-57.  
PMID: 8276818 [PubMed - indexed for MEDLINE]
- ☐ 52: [Zhang MQ, Marr TG.](#) Related Articles, Links  
 Fission yeast gene structure and recognition.  
Nucleic Acids Res. 1994 May 11;22(9):1750-9.  
PMID: 8202381 [PubMed - indexed for MEDLINE]
- ☐ 53: [Comelli P, Konig J, Werr W.](#) Related Articles, Links  
 Alternative splicing of two leading exons partitions promoter activity between the coding regions of the maize homeobox gene Zmbox1a and Trap (transposon-associated protein).  
Plant Mol Biol. 1999 Nov;41(5):615-25.  
PMID: 10645721 [PubMed - indexed for MEDLINE]
- ☐ 54: [Sheppard M, McCoy RJ, Werner W.](#) Related Articles, Links  
 Genomic mapping and sequence analysis of the fowl adenovirus serotype 10 hexon gene.  
J Gen Virol. 1995 Oct;76 ( Pt 10):2595-600.  
PMID: 7595364 [PubMed - indexed for MEDLINE]
- ☐ 55: [Roy K, Mitsugi K, Sirotinak FM.](#) Related Articles, Links  
 Organization and alternate splicing of the murine folylpolyglutamate synthetase gene. Different splice variants in L1210 cells encode mitochondrial or cytosolic forms of the enzyme.  
J Biol Chem. 1996 Sep 27;271(39):23820-7.  
PMID: 8798611 [PubMed - indexed for MEDLINE]



☐ **56:** [Shew JY, Chen PL, Bookstein R, Lee EY, Lee WH.](#)

[Related Articles](#), [Links](#)



Deletion of a splice donor site ablates expression of the following exon and produces an unphosphorylated RB protein unable to bind SV40 T antigen.

Cell Growth Differ. 1990 Jan;1(1):17-25.

PMID: 1964074 [PubMed - indexed for MEDLINE]

☐ **57:** [Dorai T, Levy JB, Kang L, Brugge JS, Wang LH.](#)

[Related Articles](#), [Links](#)



Analysis of cDNAs of the proto-oncogene c-src: heterogeneity in 5' exons and possible mechanism for the genesis of the 3' end of v-src.

Mol Cell Biol. 1991 Aug;11(8):4165-76.

PMID: 1712905 [PubMed - indexed for MEDLINE]

☐ **58:** [Zhao SH, Simmons DG, Cross JC, Scheetz TE, Casavant TL, Soares MB, Tuggle CK.](#)

[Related Articles](#), [Links](#)



PLET1 (C11orf34), a highly expressed and processed novel gene in pig and mouse placenta, is transcribed but poorly spliced in human.

Genomics. 2004 Jul;84(1):114-25.

PMID: 15203209 [PubMed - indexed for MEDLINE]

☐ **59:** [Tolner B, Roy K, Sirotinak FM.](#)

[Related Articles](#), [Links](#)



Structural analysis of the human RFC-1 gene encoding a folate transporter reveals multiple promoters and alternatively spliced transcripts with 5' end heterogeneity.

Gene. 1998 May 12;211(2):331-41.

PMID: 9602167 [PubMed - indexed for MEDLINE]

☐ **60:** [Zhu HQ, Hu GQ, Ouyang ZQ, Wang J, She ZS.](#)

[Related Articles](#), [Links](#)



Accuracy improvement for identifying translation initiation sites in microbial genomes.

Bioinformatics. 2004 Dec 12;20(18):3308-17. Epub 2004 Jul 9.

PMID: 15247104 [PubMed - indexed for MEDLINE]

Items 41 - 60 of 100

[Previous](#)

Page

3

of 5 [Next](#)

Display

Summary

Show

20

Sort by

Send to

[Write to the Help Desk](#)

[NCBI](#) | [NLM](#) | [NIH](#)

[Department of Health & Human Services](#)

[Privacy Statement](#) | [Freedom of Information Act](#) | [Disclaimer](#)

Aug 14 2006 08:07:58



A service of the National Library of Medicine  
and the National Institutes of Health

My NCBI

[\[Sign In\]](#) [\[Register\]](#)

All Databases

PubMed

Nucleotide

Protein

Genome

Structure

OMIM

PMC

Journals

Book

Search  for

Limits

Preview/Index

History

Clipboard

Details

Display  Show  Sort by  Send to

About Entrez

Text Version

All: 100 Review: 0 ☐

Items 61 - 80 of 100

Previous  4 of 5 Next

Entrez PubMed

Overview

Help | FAQ

Tutorials

New/Noteworthy E-Utilities

PubMed Services

Journals Database

MeSH Database

Single Citation Matcher

Batch Citation Matcher

Clinical Queries

Special Queries

LinkOut

My NCBI

Related Resources

Order Documents

NLM Mobile

NLM Catalog

NLM Gateway

TOXNET

Consumer Health

Clinical Alerts

ClinicalTrials.gov

PubMed Central

☐ **61:** [Ouyang Z, Zhu H, Wang J, She ZS.](#)

[Related Articles](#), [Links](#)



Multivariate entropy distance method for prokaryotic gene identification.  
J Bioinform Comput Biol. 2004 Jun;2(2):353-73.  
PMID: 15297987 [PubMed - indexed for MEDLINE]

☐ **62:** [Duax WL, Huether R, Pletnev VZ, Langs D, Addlagatta A, Connare S, Habegger L, Gill J.](#)

[Related Articles](#), [Links](#)



Rational genomics I: antisense open reading frames and codon bias in short-chain oxido reductase enzymes and the evolution of the genetic code.  
Proteins. 2005 Dec 1;61(4):900-6.  
PMID: 16245321 [PubMed - indexed for MEDLINE]

☐ **63:** [Laub MT, Smith DW.](#)

[Related Articles](#), [Links](#)



Finding intron/exon splice junctions using INFO, Interruption Finder and Organizer.  
J Comput Biol. 1998 Summer;5(2):307-21.  
PMID: 9672834 [PubMed - indexed for MEDLINE]

☐ **64:** [Smith DJ.](#)

[Related Articles](#), [Links](#)



Mini-exon epitope tagging for analysis of the protein coding potential of genomic sequence.  
Biotechniques. 1997 Jul;23(1):116-20.  
PMID: 9232241 [PubMed - indexed for MEDLINE]

☐ **65:** [Thanaraj TA, Robinson AJ.](#)

[Related Articles](#), [Links](#)



Prediction of exact boundaries of exons.  
Brief Bioinform. 2000 Nov;1(4):343-56.  
PMID: 11465052 [PubMed - indexed for MEDLINE]

☐ **66:** [Yada T, Takagi T, Totoki Y, Sakaki Y, Takaeda Y.](#)

[Related Articles](#), [Links](#)



DIGIT: a novel gene finding program by combining gene-finders.  
Pac Symp Biocomput. 2003;:375-87.  
PMID: 12603043 [PubMed - indexed for MEDLINE]

☐ **67:** [Strelets VB, Lim HA.](#)

[Related Articles](#), [Links](#)



Ancient splice junction shadows with relation to blocks in protein structure.  
Biosystems. 1995;36(1):37-41.

PMID: 8527694 [PubMed - indexed for MEDLINE]

- ☐ **68:** [Haviland DL, Haviland JC, Fleischer DT, Wetsel RA.](#) [Related Articles, Links](#)



Structure of the murine fifth complement component (C5) gene. A large, highly interrupted gene with a variant donor splice site and organizational homology with the third and fourth complement component genes.

J Biol Chem. 1991 Jun 25;266(18):11818-25.

PMID: 1711041 [PubMed - indexed for MEDLINE]

- ☐ **69:** [Bateman JF, Chan D, Moeller I, Hannagan M, Cole WG.](#) [Related Articles, Links](#)



A 5' splice site mutation affecting the pre-mRNA splicing of two upstream exons in the collagen COL1A1 gene. Exon 8 skipping and altered definition of exon 7 generates truncated pro alpha 1(I) chains with a non-collagenous insertion destabilizing the triple helix.

Biochem J. 1994 Sep 15;302 ( Pt 3):729-35.

PMID: 7945197 [PubMed - indexed for MEDLINE]

- ☐ **70:** [Silke J.](#) [Related Articles, Links](#)



The majority of long non-stop reading frames on the antisense strand can be explained by biased codon usage.

Gene. 1997 Jul 18;194(1):143-55.

PMID: 9266684 [PubMed - indexed for MEDLINE]

- ☐ **71:** [Mohamed MK, Taylor RE, Feinstein DS, Huang X, Pittler SJ.](#) [Related Articles, Links](#)



Structure and upstream region characterization of the human gene encoding rod photoreceptor cGMP phosphodiesterase alpha-subunit.

J Mol Neurosci. 1998 Jun;10(3):235-50.

PMID: 9770645 [PubMed - indexed for MEDLINE]

- ☐ **72:** [Solovyev V, Salamov A.](#) [Related Articles, Links](#)



The Gene-Finder computer tools for analysis of human and model organisms genome sequences.

Proc Int Conf Intell Syst Mol Biol. 1997;5:294-302.

PMID: 9322052 [PubMed - indexed for MEDLINE]

- ☐ **73:** [Burset M, Seledtsov IA, Solovyev VV.](#) [Related Articles, Links](#)



Analysis of canonical and non-canonical splice sites in mammalian genomes.

Nucleic Acids Res. 2000 Nov 1;28(21):4364-75.

PMID: 11058137 [PubMed - indexed for MEDLINE]

- ☐ **74:** [Stohr H, Marquardt A, White K, Weber BH.](#) [Related Articles, Links](#)



cDNA cloning and genomic structure of a novel gene (C11orf9) localized to chromosome 11q12-->q13.1 which encodes a highly conserved, potential membrane-associated protein.

Cytogenet Cell Genet. 2000;88(3-4):211-6.

PMID: 10828591 [PubMed - indexed for MEDLINE]

- ☐ **75:** [Huang JP, Tang CJ, Kou GH, Marchesi VT, Benz EJ Jr, Tang TK.](#) [Related Articles, Links](#)



Genomic structure of the locus encoding protein 4.1. Structural basis for complex combinational patterns of tissue-specific alternative RNA splicing.

J Biol Chem. 1993 Feb 15;268(5):3758-66.

PMID: 8429050 [PubMed - indexed for MEDLINE]

- ☐ **76:** [Beroud C, Carrie A, Beldjord C, Deburgrave N, Llense S, Carelle N, Peccate C, Cuisset JM, Pandit F, Carre-Pigeon F, Mayer M, Bellance R, Recan D, Chelly J, Kaplan JC, Leturcq F.](#) [Related Articles, Links](#)



Dystrophinopathy caused by mid-intronic substitutions activating cryptic exons in the DMD gene.

Neuromuscul Disord. 2004 Jan;14(1):10-8.

PMID: 14659407 [PubMed - indexed for MEDLINE]

- ☐ **77:** [Nakata K, Kanehisa M, DeLisi C.](#)

[Related Articles, Links](#)



Prediction of splice junctions in mRNA sequences.

Nucleic Acids Res. 1985 Jul 25;13(14):5327-40.

PMID: 4022782 [PubMed - indexed for MEDLINE]

- ☐ **78:** [Bussow K, Hoffmann S, Sievert V.](#)

[Related Articles, Links](#)



ORFer--retrieval of protein sequences and open reading frames from GenBank and storage into relational databases or text files.

BMC Bioinformatics. 2002 Dec 19;3:40. Epub 2002 Dec 19.

PMID: 12493080 [PubMed - indexed for MEDLINE]

- ☐ **79:** [Smith GL, Chan YS, Howard ST.](#)

[Related Articles, Links](#)



Nucleotide sequence of 42 kbp of vaccinia virus strain WR from near the right inverted terminal repeat.

J Gen Virol. 1991 Jun;72 ( Pt 6):1349-76.

PMID: 2045793 [PubMed - indexed for MEDLINE]

- ☐ **80:** [Yuasa K, Kanoh Y, Okumura K, Omori K.](#)

[Related Articles, Links](#)



Genomic organization of the human phosphodiesterase PDE11A gene.

Evolutionary relatedness with other PDEs containing GAF domains.

Eur J Biochem. 2001 Jan;268(1):168-78.

PMID: 11121118 [PubMed - indexed for MEDLINE]

Items 61 - 80 of 100

Previous **Page**  of 5 Next

Display  Show  Sort by  Send to

[Write to the Help Desk](#)

[NCBI](#) | [NLM](#) | [NIH](#)

[Department of Health & Human Services](#)

[Privacy Statement](#) | [Freedom of Information Act](#) | [Disclaimer](#)

Aug 14 2006 08:07:58



Search  for

Limits Preview/Index History Clipboard Details

Display  Show  Sort by  Send to

About Entrez

Text Version

All: 100 Review: 0

Items 81 - 100 of 100


Previous  5 of 5

Entrez PubMed

Overview

Help | FAQ

Tutorials

New/Noteworthy   
E-Utilities

PubMed Services

Journals Database

MeSH Database

Single Citation Matcher

Batch Citation Matcher

Clinical Queries

Special Queries

LinkOut

My NCBI

Related Resources

Order Documents

NLM Mobile

NLM Catalog

NLM Gateway

TOXNET

Consumer Health

Clinical Alerts

ClinicalTrials.gov

PubMed Central

☐ **81:** [Gotlib RW, Bishop DF, Wang AM, Zeidner KM, Ioannou YA, Adler DA, Distech CM, Desnick RJ.](#) Related Articles, Links



The entire genomic sequence and cDNA expression of mouse alpha-galactosidase A.

Biochem Mol Med. 1996 Apr;57(2):139-48.

PMID: 8733892 [PubMed - indexed for MEDLINE]

☐ **82:** [Burset M, Guigo R.](#) Related Articles, Links



Evaluation of gene structure prediction programs.

Genomics. 1996 Jun 15;34(3):353-67.

PMID: 8786136 [PubMed - indexed for MEDLINE]

☐ **83:** [Zorio DA, Lea K, Blumenthal T.](#) Related Articles, Links



Cloning of Caenorhabditis U2AF65: an alternatively spliced RNA containing a novel exon.

Mol Cell Biol. 1997 Feb;17(2):946-53.

PMID: 9001248 [PubMed - indexed for MEDLINE]

☐ **84:** [Claverie JM.](#) Related Articles, Links



Exon detection by similarity searches.

Methods Mol Biol. 1997;68:283-313. No abstract available.

PMID: 9055264 [PubMed - indexed for MEDLINE]

☐ **85:** [Newton DC, Bevan SC, Choi S, Robb GB, Millar A, Wang Y, Marsden PA.](#) Related Articles, Links



Translational regulation of human neuronal nitric-oxide synthase by an alternatively spliced 5'-untranslated region leader exon.

J Biol Chem. 2003 Jan 3;278(1):636-44. Epub 2002 Oct 25.

PMID: 12403769 [PubMed - indexed for MEDLINE]

☐ **86:** [Snyder EE, Stormo GD.](#) Related Articles, Links



Identification of protein coding regions in genomic DNA.

J Mol Biol. 1995 Apr 21;248(1):1-18.

PMID: 7731036 [PubMed - indexed for MEDLINE]











☐ **87:** [Guigo R, Knudsen S, Drake N, Smith T.](#) Related Articles, Links




Prediction of gene structure.

J Mol Biol. 1992 Jul 5;226(1):141-57.


PMID: 1619647 [PubMed - indexed for MEDLINE]

- ☐ **88:** [Vignal L, d'Aubenton-Carafa Y, Lisacek F, Mephu Nguifo E, Rouze P, Quinqueton J, Thermes C.](#) Related Articles, Links  
 Exon prediction in eucaryotic genomes.  
Biochimie. 1996;78(5):327-34.  
PMID: 8905152 [PubMed - indexed for MEDLINE]
- ☐ **89:** [Guigo R.](#) Related Articles, Links  
 Computational gene identification: an open problem.  
Comput Chem. 1997;21(4):215-22.  
PMID: 9415986 [PubMed - indexed for MEDLINE]
- ☐ **90:** [Li W.](#) Related Articles, Links  
 Statistical properties of open reading frames in complete genome sequences.  
Comput Chem. 1999 Jun 15;23(3-4):283-301.  
PMID: 10404621 [PubMed - indexed for MEDLINE]
- ☐ **91:** [Campione-Piccardo J.](#) Related Articles, Links  
 Fast algorithms for predicting operational reading frames following splicing or insertion in expression vectors.  
Comput Biomed Res. 1987 Oct;20(5):405-9.  
PMID: 2824122 [PubMed - indexed for MEDLINE]
- ☐ **92:** [Lopez R, Larsen F, Prydz H.](#) Related Articles, Links  
 Evaluation of the exon predictions of the GRAIL software.  
Genomics. 1994 Nov 1;24(1):133-6.  
PMID: 7896267 [PubMed - indexed for MEDLINE]
- ☐ **93:** [Kent C, Landau GM, Ziv-Ukelson M.](#) Related Articles, Links  
 On the complexity of sparse exon assembly.  
J Comput Biol. 2006 Jun;13(5):1013-27.  
PMID: 16796548 [PubMed - indexed for MEDLINE]
- ☐ **94:** [Cai Y, Bork P.](#) Related Articles, Links  
 Homology-based gene prediction using neural nets.  
Anal Biochem. 1998 Dec 15;265(2):269-74.  
PMID: 9882402 [PubMed - indexed for MEDLINE]
- ☐ **95:** [Kleffe J, Hermann K, Borodovsky M.](#) Related Articles, Links  
 Statistical analysis of GeneMark performance by cross-validation.  
Comput Chem. 1996 Mar;20(1):123-33.  
PMID: 16749185 [PubMed - indexed for MEDLINE]
- ☐ **96:** [Fickett JW.](#) Related Articles, Links  
 The gene identification problem: an overview for developers.  
Comput Chem. 1996 Mar;20(1):103-18.  
PMID: 16749184 [PubMed - indexed for MEDLINE]
- ☐ **97:** [Gaedigk A, Leeder JS.](#) Related Articles, Links  
 Comments on Hoskins et al. [(2005) drug metab dispos 33:1564-1565].  
Drug Metab Dispos. 2006 Mar;34(3):504-5; author reply 506. No abstract available.  
PMID: 16495382 [PubMed - indexed for MEDLINE]
- ☐ **98:** [Huang J, Li T, Chen K, Wu J.](#) Related Articles, Links

-  **An approach of encoding for prediction of splice sites using SVM.**  
Biochimie. 2006 Jul;88(7):923-929. Epub 2006 Apr 3.  
PMID: 16626852 [PubMed - as supplied by publisher]


☐ **99:** [Cebrat S, Dudek MR.](#)

[Related Articles, Links](#)

-  **Generation of overlapping open reading frames.**  
Trends Genet. 1996 Jan;12(1):12. No abstract available.  
PMID: 8741854 [PubMed - indexed for MEDLINE]





☐ **100:** [Sopko R, Andrews B.](#)

[Related Articles, Links](#)

-  **Small open reading frames: not so small anymore.**  
Genome Res. 2006 Mar;16(3):314-5. No abstract available.  
PMID: 16510897 [PubMed - in process]

Items 81 - 100 of 100

Previous **Page**  of 5

Display   Show   Sort by   Send to  

[Write to the Help Desk](#)  
[NCBI](#) | [NLM](#) | [NIH](#)  
[Department of Health & Human Services](#)  
[Privacy Statement](#) | [Freedom of Information Act](#) | [Disclaimer](#)

Aug 14 2006 08:07:58



A service of the National Library of Medicine  
and the National Institutes of Health

My NCBI  
[Sign In] [Reg]

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Bc

Search PubMed for [ ] Preview Go C

Limits Preview/Index **History** Clipboard Details

About Entrez

Text Version

- Search History will be lost after eight hours of inactivity.
- To combine searches use # before search number, e.g., #2 AND #6.
- Search numbers may not be continuous; all searches are represented.
- Click on query # to add to strategy

Entrez PubMed

Overview

Help | FAQ

Tutorials

New/Noteworthy

E-Utilities

PubMed Services

Journals Database

MeSH Database

Single Citation Matcher

Batch Citation Matcher

Clinical Queries

Special Queries

LinkOut

My NCBI

Search	Most Recent Queries	Time	Result
<a href="#">#15</a>	<b>Related Articles for PubMed (Select 1321415)</b>	12:49:38	<a href="#">100</a>
<a href="#">#10</a>	<b>Related Articles for PubMed (Select 7584412)</b>	12:48:00	<a href="#">102</a>
<a href="#">#7</a>	<b>Search nucleotide and sequence and analysis and method and atg</b>	12:34:42	<a href="#">35</a>
<a href="#">#5</a>	<b>Search nucleotide and sequence and analysis and method</b>	12:33:56	<a href="#">8285</a>
<a href="#">#4</a>	<b>Search predict and protein and dna and atg</b>	12:31:58	<a href="#">9</a>
<a href="#">#2</a>	<b>Search dna and sequence and met and start and stop</b>	12:29:20	<a href="#">9</a>
<a href="#">#1</a>	<b>Search dna sequence analysis and met</b>	12:28:33	<a href="#">416</a>

Clear History

Related Resources

Order Documents

NLM Mobile

NLM Catalog

NLM Gateway

TOXNET

Consumer Health

Clinical Alerts

ClinicalTrials.gov

PubMed Central

[Write to the Help Desk](#)

[NCBI](#) | [NLM](#) | [NIH](#)

[Department of Health & Human Services](#)

[Privacy Statement](#) | [Freedom of Information Act](#) | [Disclaimer](#)

Aug 14 2006 08:07:58





A service of the National Library of Medicine  
and the National Institutes of Health

My NCBI  
[Sign In] [Register]

All Databases

PubMed

Nucleotide

Protein

Genome

Structure

OMIM

PMC

Journals

Books

Search PubMed for [ ] Go Clear

Limits

Preview/Index

History

Clipboard

Details

Display AbstractPlus Show 20 Sort by Send to

About Entrez

Text Version

Entrez PubMed

Overview

Help | FAQ

Tutorials

New/Noteworthy

E-Utilities

PubMed Services

Journals Database

MeSH Database

Single Citation Matcher

Batch Citation Matcher

Clinical Queries

Special Queries

LinkOut

My NCBI

Related Resources

Order Documents

NLM Mobile

NLM Catalog

NLM Gateway

TOXNET

Consumer Health

Clinical Alerts

ClinicalTrials.gov

PubMed Central

☐ 1: Nucleic Acids Res. 1982 Sep 11;10(17):5303-18.

**FREE full text article**  
in PubMed Central

Link:

### Recognition of protein coding regions in DNA sequences.

#### Fickett JW.

We give a test for protein coding regions which is based on simple and universal differences between protein-coding and noncoding DNA. The test is simple enough to use without a computer and is completely objective. The test has been thoroughly proven on 400,000 bases of sequence data: it misclassifies 5% of the regions tested and gives an answer of "No Opinion" one fifth of the time. We predict some new coding and noncoding regions in published sequences.

PMID: 7145702 [PubMed - indexed for MEDLINE]

### Related Links

Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach [J Mol Biol. 1991]

Average mutual information of coding and noncoding DNA [Proc Natl Acad Sci USA. 2000]

Locating protein coding regions in human DNA using a decision tree algorithm [J Comput Biol. 1995]

[Analysis, identification and correction of some errors of model refseqs appeared in NCBI Human Gene Database by in silico cloning and experimental verification of novel human genes] [Yi Chuan Xue Bao. 2004]

Detection of protein coding sequences using a mixture model for local protein amino acid composition [Comput Biol. 2000]

See all Related Articles...

Display AbstractPlus Show 20 Sort by Send to

Write to the Help Desk

NCBI | NLM | NIH

Department of Health & Human Services

Privacy Statement | Freedom of Information Act | Disclaimer

Aug 14 2006 08:07:58

Recognition of protein coding regions in DNA sequences

James W. Fickett

Theoretical Division, Los Alamos National Laboratory, University of California, Los Alamos, NM 87544, USA

Received 6 July 1982; Revised and Accepted 20 July 1982

ABSTRACT

We give a test for protein coding regions which is based on simple and universal differences between protein-coding and noncoding DNA. The test is simple enough to use without a computer and is completely objective. The test has been thoroughly proven on 400,000 bases of sequence data: it misclassifies 5% of the regions tested and gives an answer of "No Opinion" one fifth of the time. We predict some new coding and noncoding regions in published sequences.

INTRODUCTION

There has been for several years now a well known and very general need for a way to distinguish a true protein-coding sequence (PCS) from a merely fortuitous open reading frame (ORF) in known DNA sequences. The need arises mainly when a gene location is only approximately known at the start of sequencing, and the sequence turns out to have more than one candidate ORF. Even when a gene has been located the surrounding sequence may contain other ORF's of unknown character, and a method to distinguish the true PCS's among these yields a powerful tool for the discovery and characterization of new proteins.

We set ourselves the task of finding an objective and self-contained test, (or decision procedure) which when presented with a DNA sequence would classify it as either coding or noncoding (in this paper "coding" will always mean "coding for protein"). Later we decided to allow the test the option of refusing to classify an occasional sequence. To be of practical value such a test should not depend on the subjective evaluation of results by the user, and should have been checked on a large number of sequences so as to be of known reliability. We chose to look for a test depending on the overall statistical properties of the base sequence rather than on specific transcription or translation initiation signals for two reasons. First, initiation signals may be unavailable. This happens frequently when the 5'

end of an interesting ORF is not included in the known sequence. It can also happen that a PCS has no initiation signals at all: cf. for example the lysis gene of phage MS2, which is only translated upon readthrough of the stop codon of the previous gene (1), and the yeast mitochondrial introns which code for protein (reviewed in Ref. 2). Second, the problem of precisely characterizing what is and what is not an initiation signal still looks extremely difficult. We also chose to find a test which would give a simple coding/noncoding answer for a specific region, rather than trying to map all coding and noncoding regions in a large sequence at once. This makes it easier to do meaningful large-scale reliability testing. Also, though our test is not adapted to finding the exact boundaries of coding regions, it is very well adapted for combination with other relevant algorithms, such as searches for ORF's, ribosome binding sites, intron boundaries, etc.

Four papers have appeared in the last year which describe statistical patterns which are probably characteristic of coding regions in general. All of these patterns have the potential of forming the basis for a useful coding/noncoding test. However we believe that ours is the first paper to give a fully specified and objective test, checked on a large number of sequences. Shulman et al. found (3) patterns in the coding regions of two phage that pointed to the three letter code and to the correct reading frame. However their sample was very small, and they did not investigate the predictive power of their observations. J. C. W. Shepherd, in researching the origin of the genetic code, found (4) periodicities in the autocorrelation functions of single bases and doublets in DNA, and applied this (5) to the problem of discovering the reading frame of a PCS. Though interesting patterns are found, no specific coding/noncoding test is given, and no evidence is presented that noncoding DNA always lacks the patterns supposedly characteristic of coding DNA. Staden and McLachlan have written (6) a computer program for mapping the PCS's in a sequence by measuring the similarity of the codon usage strategy between a known PCS and the ORF under test. The method requires that the PCS used as a standard be closely related (in codon usage patterns) to any PCS discovered. This makes the method highly dependent on the judgement of the user, and may make it inapplicable in some cases.

Another, more popular, vein of research is in trying to characterize the signals for initiation of transcription and translation by which the cell itself recognizes a PCS. For reasons given above we consider this a separate problem, complementary to the one we are considering, and only refer the

reader to the surveys of Gold et al. (7) and Breathnach and Chambon (8), and to the recent computer program of Rodier et al. (9).

#### CHARACTERISTIC PARAMETERS OF CODING AND NONCODING REGIONS

Many people have noticed patterns, or statistical order, in PCS's, but for the most part it has not been shown that these patterns consistently fail to appear in noncoding DNA. In this section we will give a striking illustration to show that some of the order in PCS's is in fact characteristic of coding regions, and will then define some numerical parameters of sequences whose distributions reveal universal differences between coding and noncoding DNA.

All studies reported here are based on sequence data stored in the Los Alamos Sequence Library, a public databank on the CDC 7600 computers at Los Alamos National Laboratory, currently listing 486,000 bases in 320 sequences. A description of the databank (including references for the sequences) is given in Ref. 10. Each sequence in the library was divided into its coding and noncoding parts, based on the experimental evidence reported by the original authors: sections of sequence for which this information was incomplete were not used. In early experiments we found that sequences under 200 bases (a somewhat arbitrary limit, considered further below) were too small to give reliable results. So for our primary data we took 321 fragments of coding DNA (230877 bases) and 249 fragments of noncoding DNA (158987 bases), each at least 200 bases long. (Thus a coding/noncoding decision made by the test given in this paper is based on the data in the Los Alamos Sequence Library. But we will show that our method is general and can be based on any collection of sequence data.)

Underlying all observations of statistical order in PCS's is the fact that codons are used with unequal frequency (for data and review see the work of Grantham et al. (11-13)). One consequence of this fact, which has been noted several times (3-5,14,15), is that oligonucleotides (and in particular nucleotides) tend to be repeated with a periodicity of three in a PCS. Figure 1 shows the autocorrelation function for thymine in the coding and noncoding parts of the Los Alamos Library (we ignore the distinction between RNA and DNA throughout the paper, so T and U are considered synonymous). The first graph shows that in coding sequences the number of bases separating two T's is much more likely to be 2,5,8,11,... ( $2+3n$ ) than it is to be  $3n$  or  $1+3n$ . I.e. in coding sequences identical bases are most often found in identical codon positions. The second graph shows that this regularity is absent in noncoding sequences.

We now turn to the definition of eight numerical parameters of DNA sequences which we use to distinguish coding from noncoding regions. The first four parameters, motivated by Figure 1, measure the asymmetry in the distribution of each base among the three codon positions (or the analogous positions in a noncoding sequence). Let

- $A_1$  = Number of A's in positions 1,4,7,10,...  
 (1)  $A_2$  = Number of A's in positions 2,5,8,11,...  
 $A_3$  = Number of A's in positions 3,6,9,12,...

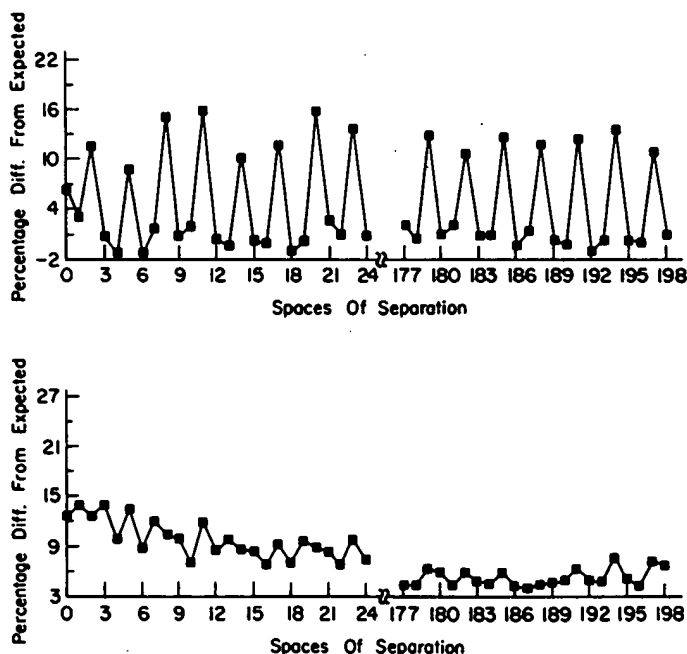


FIGURE 1. Autocorrelation graphs for T (thymine) in the 321 coding and 249 noncoding fragments over 200 bases long in the Los Alamos Sequence Library. Top: For each possible separation  $k$ , we counted, in all coding fragments in the Library (a total of 231 kilobases) the number of times two T's appeared with  $k$  nucleotides between them, and compared this with the count expected in a model where bases are chosen independently - namely the number of blocks of  $k+2$  nucleotides times the square of the overall T-content of the coding regions. The percent difference is graphed for  $k$  running from 0 to 24 and from 147 to 198. Bottom: The same for the noncoding regions (159 kilobases). The wave so conspicuous for the coding regions is absent here. Findings were similar for the other three bases, and for pairs of unlike bases. The high values near the beginning of the noncoding graph are probably due to AT clustering; otherwise the two graphs have about the same average value.

and similarly for C, G and T. Then define

$$(2) \quad A\text{-Position} = \frac{\text{MAX}(A_1, A_2, A_3)}{\text{MIN}(A_1, A_2, A_3)+1}$$

and similarly for C, G and T.

The parameters A-, C-, G- and T-Position measure the degree to which each base is favored in one codon position over another. Note that it is irrelevant which of the three codon positions favors the base; it is only the degree to which the base is favored that is measured - this property gives these four parameters fairly similar distributions in all sequences, regardless of the well known differences in codon usage strategy between organisms.

The other four parameters we use are just the A-, C-, G- and T-Content of the sequence (i.e. the percentage of the sequence contributed by each of four bases). Note that, as a practical matter, the counts  $A_1$  etc. made in the calculation of the Position parameters yield immediately the Content parameters also.

The relative distribution of these eight parameters between coding and noncoding fragments is shown in Table 1. All eight parameters will be used in a single test in the next section, but note that even in the distribution of individual parameters the differences between coding and noncoding DNA are evident. For example among fragments having a T-Position parameter less than 1.2 (this includes about one fourth of all fragments) there is only a 9% probability of coding function, while among fragments with T-Position parameter over 1.7 (again about one fourth of the total) the probability of coding is over 90%. Table 1 contains all the information about these parameters needed for our decision procedure. The full distributions of the eight parameters, of interest in their own right, are given in Figure 2 and discussed further below.

#### HOW TO DISTINGUISH CODING FROM NONCODING SEQUENCES

In the last section we gave the distribution of our eight test parameters. Next we will assign weights to each parameter, telling how much attention we should pay to it in making the final coding/noncoding decision. The parameter distribution and weights should need to be recalculated only very occasionally as more sequence data accumulates. Users of the coding/noncoding test will only need to do a very simple calculation detailed below.

From Table 1 it is clear that, for example, the T-Position parameter of a sequence usually tells one a good deal more than its A-Content. To get a

TABLE 1  
Characteristic Parameters of Coding and Noncoding Sequences

<u>Position Parameter</u>		<u>Probability of Coding</u>			
0.0 to 1.1		A: .22	C: .23	G: .08	T: .09
1.1 1.2		.20	.30	.08	.09
1.2 1.3		.34	.33	.16	.20
1.3 1.4		.45	.51	.27	.54
1.4 1.5		.68	.48	.48	.44
1.5 1.6		.58	.66	.53	.69
1.6 1.7		.93	.81	.64	.68
1.7 1.8		.84	.70	.74	.91
1.8 1.9		.68	.70	.88	.97
1.9 2.0+		.94	.80	.90	.97

<u>Content Parameter</u>		<u>Probability of Coding</u>			
.00 to .17		A: .21	C: .31	G: .29	T: .58
.17 .19		.81	.39	.33	.51
.19 .21		.65	.44	.41	.69
.21 .23		.67	.43	.41	.56
.23 .25		.49	.59	.73	.75
.25 .27		.62	.59	.64	.55
.27 .29		.55	.64	.64	.40
.29 .31		.44	.51	.47	.39
.31 .33		.49	.64	.54	.24
.33 .99		.28	.82	.40	.28

TABLE 1. The values of the eight parameters, A-, C-, G- and T-Position and A-, C-, G- and T-Content, were calculated for each of the 321 coding and 249 noncoding fragments over 200 bases long in the Los Alamos Sequence Library (see text). The range of each parameter was divided into ten intervals as shown (we use these same intervals for any collection of sequence data). For each interval the percentage of coding and noncoding fragments whose parameter fell therein was recorded. The value "Probability of Coding" shown is the percentage of coding fragments falling in the interval, divided by the percentage of coding plus the percentage of noncoding. This is essentially the fraction of all fragments falling in the interval which are coding, but differs slightly because more coding than noncoding fragments are used.

number telling us how much input each parameter should have in the final decision, we used each parameter alone to predict coding function, as follows: if a sequence fell in an interval where the probability of coding (from Table 1) was greater than one half the sequence was called coding, otherwise not. (I.e. if more coding than noncoding fragments share this parameter value with the fragment in question, we guess it is coding.) The weight for a given parameter is just the percentage of the time that this guess was correct, less 50% (random level). The weights for each of the eight parameters are shown in Table 2. In giving these weights we are not making any important claim about

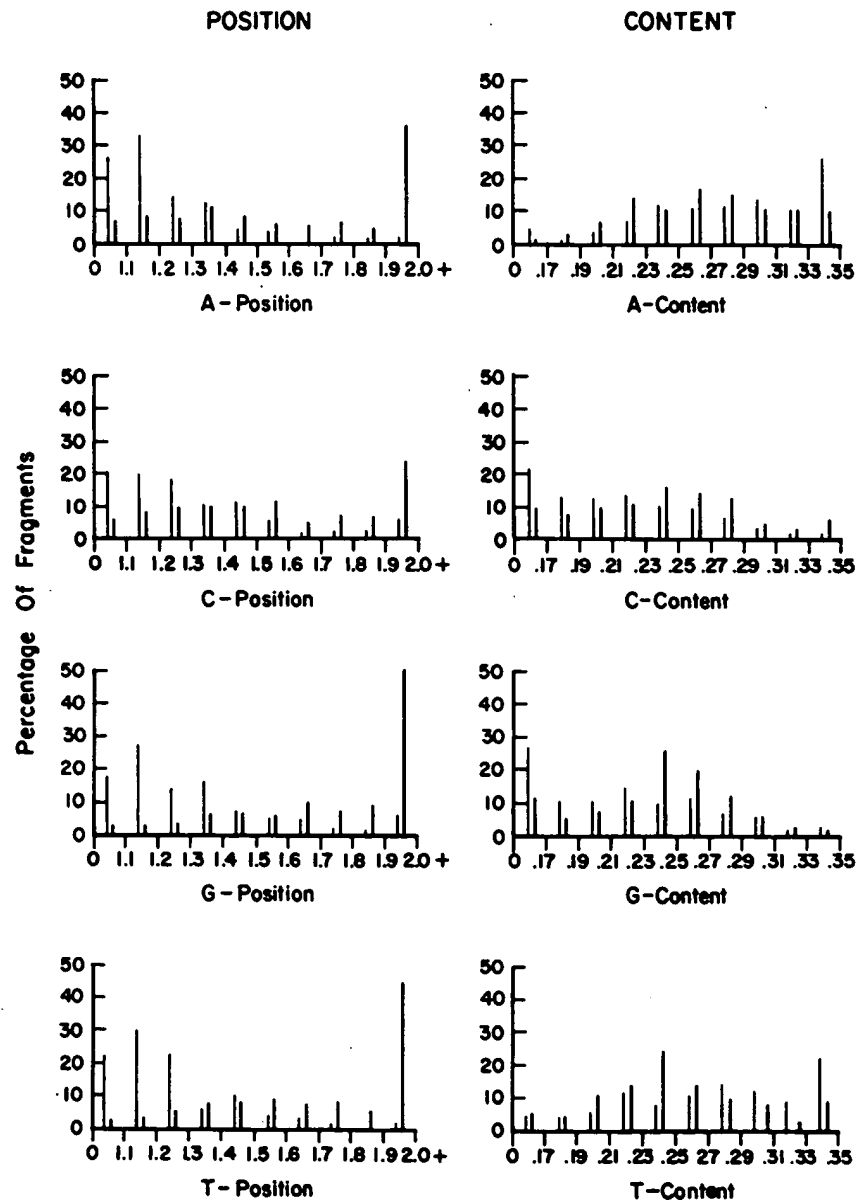


FIGURE 2. The distribution of the Position and Content parameters for coding (heavy bars) and noncoding (light bars) fragments. See the legend of Table 1 for details.



TABLE 2  
Weight to be Given to the Individual Parameters

	<u>Position</u>	<u>Content</u>
A	.26	.11
C	.18	.12
G	.31	.15
T	.33	.14

TABLE 2. The weight shown is the percentage of the time (above 50%, the random level) that each parameter alone successfully predicted coding or noncoding function.

these parameters; rather we are just deciding how to use them in our specific decision procedure.

We can now describe TESTCODE, our algorithm for predicting whether a fragment of DNA is coding or not. Given a fragment of DNA, first make the counts  $A_i$ ,  $C_i$ ,  $G_i$  and  $T_i$ ,  $i=1,3$  (equation (1)). From these calculate the eight parameters A-, C-, G- and T-Position (equation (2)), and A-, C-, G- and T-Content of the fragment. For each of these parameters look up the "Probability of Coding" value in Table 1; call these probabilities  $p_1, \dots, p_8$ . Let the corresponding weights, given in Table 2, be denoted  $w_1, \dots, w_8$ . The sum  $p_1 w_1 + \dots + p_8 w_8$  is the TESTCODE indicator of coding function. Its distribution in the Los Alamos Library, and the predictions corresponding to its different values, are shown in Table 3. (A more familiar way to combine the information from the eight parameters would be to use Bayes' formula. But in using Bayes' formula we assume that the eight parameters are independent, which of course is not the case. So it is not surprising that the method given above worked a little better.)

#### RELIABILITY OF THE METHOD

From Table 3 it is clear that TESTCODE correctly predicted the function of all but a few of the fragments used in the study. However since we used these same fragments to calculate the parameter distributions which TESTCODE uses, one might object that perhaps the algorithm was just "remembering" special properties of the Los Alamos collection, and would be less reliable for distinguishing coding and noncoding DNA in general. To take care of this objection we divided the Los Alamos Library into two parts, calculated the distribution of our eight parameters on one half, and used this information to predict which fragments in the other half coded for protein. There was only a

TABLE 3  
Distribution of the TESTCODE Indicator

<u>TESTCODE Indicator</u>	<u>Probability of Coding</u>	<u>Prediction</u>
0.32 to 0.43	0.00	Noncoding
0.43 to 0.53	0.04	Noncoding
0.53 to 0.64	0.07	Noncoding
0.64 to 0.74	0.29	Noncoding
0.74 to 0.84	0.40	No Opinion
0.84 to 0.95	0.77	No Opinion
0.95 to 1.05	0.92	Coding
1.05 to 1.16	0.98	Coding
1.16 to 1.26	1.00	Coding
1.26 to 1.37	1.00	Coding

TABLE 3. The distribution of the TESTCODE indicator, our predictor of coding function, is shown on all the 321 coding and 249 noncoding fragments used in this study. "Probability of Coding" is calculated just as in Table 1. The last column gives the TESTCODE prediction of function for a fragment whose indicator value falls in the corresponding interval. In calibrating TESTCODE on any set of sequence data there is always a natural cutoff point (in this case .84) above which every interval contains more coding than noncoding fragments, and below which every interval contains more noncoding than coding fragments. We always make the two intervals flanking this cutoff the "No Opinion" range.

5% error rate in these predictions, showing that TESTCODE is almost certainly based on universal differences between coding and noncoding DNA, independent of the Los Alamos collection.

In more detail our procedure was as follows: We numbered the coding fragments from 1 to 321 and the noncoding from 1 to 249. We then calculated the relative distribution of our eight parameters, as in Table 1, and the weights to use with them, as in Table 2, but using only the odd-numbered fragments as our data set. We then used the resulting parameter distributions to calculate a TESTCODE indicator for each of the even-numbered fragments. The range of the indicator was divided into 10 equal intervals, as in Table 3. Any fragment whose indicator fell in the top four intervals was judged coding, any in the bottom four noncoding, and in the middle two intervals no answer was given. The TESTCODE prediction was "No Opinion" on 18% of the fragments. 6% of the coding segments were judged incorrectly as "Noncoding", and 3% of the noncoding segments were judged incorrectly as "Coding". The actual distribution of the TESTCODE indicator is given in Figure 3.

In the future, when a larger sample of sequences is available, it may be

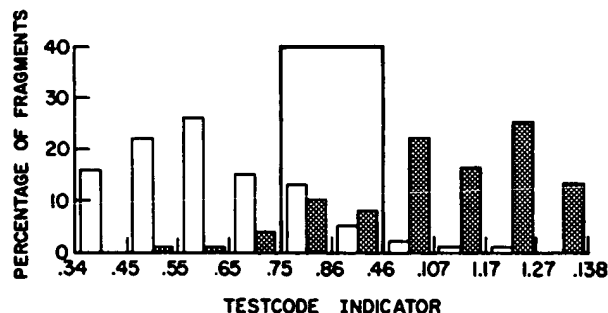


FIGURE 3. Results of the reliability test for TESTCODE. After the TESTCODE indicator was calculated for all even fragments the range of the indicator was divided into ten equal intervals, whose endpoints are marked on the abscissa. The percentage of coding (shaded bars) and noncoding (open bars) fragments whose TESTCODE indicator fell in each interval is graphed. The "No Opinion" range is boxed.

worthwhile to use separate data sets when using TESTCODE on fragments from different taxonomic classes. For example when we ran the kind of reliability test just described using only vertebrate nuclear sequences, we found that TESTCODE returned "No Opinion" on only 12% of the fragments used, and only misclassified 3%. For the vertebrate study we used 82 coding and 102 noncoding fragments; other taxonomic groups are still rather small for this kind of reliability test.

Throughout this study we have restricted attention to fragments over 200 bases long. It turns out that in fact TESTCODE's reliability is unacceptable on shorter fragments. When we used TESTCODE (just as specified in the preceeding section) to predict the function of the 57 noncoding and 159 coding fragments in the library between 100 and 199 bases in length, the predictions were incorrect 13% of the time, and the "No Opinion" rate was 29%. 200 bases seems to be a reasonable minimum, for when predictions were made in the length ranges 200-299, 300-399, 400-499, 500-599, 300+ and 600+ the error rate was always close to 5%. The chief effect of the length, above 200 bases, seems to be on the "No Opinion" rate, which is 24% for fragments of 200-299 bases, but under 15% for longer fragments.

#### PREDICTION OF CODING AND NONCODING REGIONS IN PUBLISHED SEQUENCES

We have scanned the Los Alamos Sequence Library for ORF's not associated with a known protein, and have rated them all with TESTCODE. In this section we give a few of our more interesting findings. Our predictions are summarized in Table 4; further comments on some are given below. A general

TABLE 4  
Predicted Coding and Noncoding ORF's

<u>Organism</u>	<u>Reported Sequence</u>	<u>Ref.</u>	<u>Open Frame</u>	<u>Prediction</u>
Adenovirus7	Transforming Region	(17)	402 to 166 <sup>++</sup>	Coding (.92)
A. nidulans	Cytochrome B	(18)	507 to 713 <sup>++</sup>	Coding (.98)
E. coli	Insertion Element I	(19,20)	250 to 753 <sup>#</sup> 56 to 331 <sup>#</sup>	Coding (.98) No Opinion(.77)
E. coli	Origin of Replication	(21-24)	734 to 291 <sup>++</sup> 1282 to 824 <sup>*</sup>	Coding (.92) Coding (1.0)
E. coli	Ribosomal Operon B	(25-28)	275 to 1144 <sup>#</sup> 2699 to 2959 <sup>#</sup> 6916 to 7506 <sup>#</sup>	Coding (.98) No Opinion(.77) Coding (1.0)
Human	$\delta$ -hemoglobin	(29)	1493 to 1810	Coding (.98)
Yeast	18S rRNA	(30)	1349 to 1149 <sup>*</sup>	Coding (.98)
Yeast	2 $\mu$ plasmid	(31,32)	5570 to 523 <sup>#</sup> 2008 to 887 <sup>++</sup> 5198 to 4308 <sup>++</sup> 2271 to 2816 <sup>#</sup> 6258 to 5905 <sup>++</sup>	Noncoding (.29) No Opinion(.77) Coding (.98) No Opinion(.40) Noncoding (.04)

TABLE 4. As far as we know none of the ORF's listed here has been shown to be coding or noncoding. Numbering of the sequence is as in the (first of the) reference(s) cited. The ranking by TESTCODE (from Table 3) is given in the last column.

\*Complementary strand from that given in reference.

<sup>+</sup>No start codon.

<sup>#</sup>Possible coding function suggested in reference.

experimental method for identifying the protein product of any ORF, if it exists, has been given (16), so these predictions provide a way to assess the usefulness of TESTCODE as an exploratory tool.

The gene products of the Adenovirus transforming region are of great interest, yet we have seen no mention of the Adenovirus ORF listed in Table 4. Although it has no start codon, it might be spliced with other ORF's upstream on the same strand.

It has been shown that the box3 intron of yeast cytochrome b codes for a protein maturase, and other yeast mitochondrial introns are suspected of coding (reviewed in Ref. 2). Waring et al. (18) have shown that the

situation in *Aspergillus nidulans* is similar to that in yeast; the single intron in the cytochrome b gene of *A. nidulans* has a long ORF which continues in phase with the previous exon. Since the probability, according to TESTCODE, that this ORF codes is .98, it looks very likely that coding introns will be found in organisms other than yeast.

There is considerable interest in protein products which may be coded by movable DNA elements and which may help to insert and excise them. TESTCODE ranks very highly one long ORF in Insertion Element I of *E. coli*. Ohtsubo et al. (20) have sequenced an analogue of this insertion element in *Shigella dysenteriae* and have shown that in this ORF (and another which is ranked ambiguously by TESTCODE) many more of the differences from Insertion Element I occur in third codon position than in first or second - a strong indication that both ORF's code.

The first ORF listed for the *E. coli* replication origin has been noted before, and in fact evidence that supports its probable coding function is given in Ref. 23. The second ORF listed, however, seems to have escaped attention.

The 3' flanking regions of many vertebrate genes have short ORF's, partly overlapping the gene, which rank highly. We include one fairly long one associated with Human  $\delta$ -hemoglobin, which is clearly separate from the main gene. (25% of the designated ORF overlaps the hemoglobin gene. The remaining 75% of the ORF was tested separately and found to have a .92 probability of coding.)

The possible PCS listed for Yeast 18S rRNA is particularly interesting because no PCS is known to overlap a ribosomal RNA gene. Many ribosomal RNA genes in the Los Alamos Library contain long ORF's; the second ORF listed from the *E. coli* RRNB operon is another.

We have examined all the ORF's of an important cloning vector, the yeast 2 micron plasmid, and offer our opinion on its overall coding capacity.

#### WHY TESTCODE WORKS

In this section we show that TESTCODE's success can be understood in terms of two simple facts: 1) Any kind of consistent non-random codon use results in uniformly high Position parameters, and 2) Coding sequences have higher GC-content, on average, than noncoding sequences. We begin by explaining more fully the connection between codon usage and our Position parameters.

Suppose we had an organism in which A was suppressed in third codon position, but shared first and second codon position equally with the other

three bases. Thus the probability that the first base of a codon was A would be .25, and likewise the second, but the probability that the third base was A might be only .15. Then in a PCS of length N we would have, approximately,  $A_1 = .25N$ ,  $A_2 = .25N$  and  $A_3 = .15N$ , so that the expected value of A-Position would be about  $.25/.15$ , or 1.7. Now note that if we had another organism in which third position A was favored instead of suppressed, so that the probabilities of finding an A in each of the three positions was, say .22, .25 and .35, respectively, the expected value of the A-Position parameter would be  $.35/.22 = 1.6$ , a similar value. Thus it turns out that all the very different coding strategies used by different creatures lead to the same result - Position parameter values mostly in the range 1.5 to 4.0 (whereas noncoding fragments have Position values, generally, in the range 1.0 to 1.5). As we mentioned earlier, this is what makes our one calculation applicable to all different kinds of sequences.

To take an actual example, the probabilities of finding an A in each of the three codon positions in vertebrates are .27, .31, and .15. We would predict from this an average A-Position parameter of  $.31/.15 = 2.1$ , while the actual average is 3.2. The true average is higher because the PCS's exhibit stronger codon usage preferences individually than one sees in the overall average. In the same way the predicted average C-, G-, and T-Parameter values are 1.6, 1.6 and 1.5 respectively, while the actual averages are 1.8, 1.9 and 1.9.

As one can see from Table 2, TESTCODE's decision is based mainly on the Position parameters. However the base content of the sequence shows some clear trends and does contribute a few percent to the reliability. The most noticeable trend in the base content data is that the GC-content of coding sequences tends to be higher than that of noncoding sequences.

To test whether these statistical trends really account for TESTCODE's performance, we generated artificial random "coding" and "noncoding" sequences and rated them with TESTCODE. For our synthetic "coding" sequences we generated successive codons independently and at random, with the same frequencies as genuine vertebrate sequences. (The Library as a whole does not show strong codon preference rules, so we needed to limit ourselves to a more internally consistent set of data. There is no reason to think that the choice of vertebrate instead of, say, E. coli sequences is significant.) For our "noncoding" sequences we generated successive bases independently and at random, with frequency .27 for A and T, and .23 for G and C (again the frequencies of vertebrate sequences). We generated 100 coding and 100

## Nucleic Acids Research

---

noncoding random sequences, each 600 bases long (the average length of the real coding and noncoding fragments used). TESTCODE, using the data from real sequences listed in Tables 1-3, classified only 2% of the random sequences incorrectly, and gave an answer of "No Opinion" on only 17%.

### SUMMARY AND DISCUSSION

We have used certain universal differences between protein-coding and noncoding regions to produce a simple algorithm TESTCODE which distinguishes coding from noncoding DNA with high reliability. When TESTCODE was calibrated on one half of the Los Alamos Sequence Library and then used to predict the coding or noncoding regions in the other half it gave an answer of "No Opinion" on 18% of the regions tested, and had an overall error rate of only 5%. We have used TESTCODE to predict a number of new coding and noncoding regions in published sequences.

A method for distinguishing coding from noncoding DNA has a large number of potential uses. First, after a fragment of DNA known to contain the gene for a certain protein has been isolated and sequenced, it often turns out to contain several ORF's from among which one must choose the correct one. A recent example is the search for the E. coli trpR gene by Singleton et al. (33). The authors considered three possible ORF's and discovered the correct one by mutation analysis. TESTCODE rates only the correct one as coding. Thus TESTCODE (or a related algorithm) may be able to reduce the experimental work in such cases to a single confirmatory experiment. Second, when newly sequenced DNA is found to contain an ORF of unknown function, TESTCODE may be used to decide whether it is likely to code for a new protein. This could be a powerful technique for discovering new proteins. One can even imagine the day when semi-automated sequencing of entire genomes followed by computer analysis of the results could fully catalogue the proteins of an organism. A third use for TESTCODE is in checking the accuracy of the data in computer-based sequence libraries. We discovered several errors in the Los Alamos Library with the help of TESTCODE.

We think that TESTCODE will prove to be useful both to experimentalists in their initial analysis of sequence data and to theoreticians as they learn about the differences between coding and noncoding DNA. However we do not claim to have discovered the ultimate coding/noncoding test. Indeed, the main value of this paper as we see it is that it presents one method for recognizing coding sequences which is spelled out in complete detail and has been tried out on a large collection of sequence data. Thus other people can

easily use TESTCODE and know how to interpret the results. We will gladly make available our programs and data to anyone wishing to more fully develop and test other methods. (They are available on-line to users of the Los Alamos Library. Others may request a tape by mail.)

Research on TESTCODE-like algorithms is complementary to several other lines of research. For example on the one hand TESTCODE only has a resolution of 200 bases and can not pinpoint the exact boundaries of a PCS, while on the other hand methods for recognizing signals for the initiation of transcription, initiation of translation, and intron splicing are poorly developed and require additional confirmation; thus these two methods can profitably be combined. Also, since TESTCODE is completely insensitive to phase, it can only be used to tell when a region is coding, and not what the coding frame is. This limitation can usually be overcome by combining TESTCODE with a search for ORF's, but when two ORF's overlap in different phases, another method is needed to decide which is the correct one. This can very likely be done using published methods mentioned in the introduction (3-6). Users of TESTCODE should be aware of one other point: we have not checked TESTCODE on regions of mixed coding/noncoding character. Thus it would be best to apply TESTCODE to regions that will be either fully coding or fully noncoding, for example ORF's starting at the last probable fMET codon.

There is some interesting regularity in the errors that TESTCODE makes. In coding sequences which are incorrectly classified as noncoding it often seems that some use is being made of the DNA which causes the usual codon preference rules to be overridden. For example one of two overlapped viral genes is sometimes classified as noncoding. Also, variable regions of immunoglobulin genes often are rated noncoding, presumably because the mechanism which generates diversity of these regions is stronger than whatever force encourages consistent codon preference. A very interesting example pertains to the yeast mating type loci. The four presumptive PCS's there are rated noncoding - possibly this means that some other pattern is present in this region of the DNA which is necessary to enable transposition.

#### ACKNOWLEDGEMENTS

We gratefully acknowledge the helpful criticism and encouragement of W. Beyer, M. Dembo, W. Goad, B. Goldstein, B. Nelson and the referees. This work was performed under the auspices of the U. S. Department of Energy.



REFERENCES

1. Kastelein, R.A., Remaut, E., Fiers, W. and van Duin, J. (1982) *Nature* 295, 35-41
2. Borst, P. and Grivell, L.A. (1981) *Nature* 289, 439-440
3. Shulman, M.J., Steinberg, C.M. and Westmoreland, N. (1981) *J. Theor. Biol.* 88, 409-420
4. Shepherd, J.C.W. (1981) *J. Mol. Evol.* 17, 94-102
5. Shepherd, J.C.W. (1981) *Proc. Nat. Acad. Sci. USA* 78, 1596-1600
6. Staden, R. and McLachlan, A.D. (1982) *Nucleic Acids Res.* 10, 141-156
7. Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Singer, B.S., and Stormo, G. (1981) *Ann. Rev. Microbiol.* 35, 365-403
8. Breathnach, R. and Chambon, P. (1981) *Ann. Rev. Biochem.* 50, 349-383
9. Rodier, F., Gabarro-Arpa, J., Ehrlich, R. and Reiss, C. (1982) *Nucleic Acids Res.* 10, 391-402
10. Fickett, J.W., Goad, W.B. and Kanehisa, M. (1982) Los Alamos National Laboratory Report LA-9724-MS
11. Grantham, R., Gautier, C. and Gouy, M. (1980) *Nucleic Acids Res.* 8, 1893-1912
12. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) *Nucleic Acids Res.* 9, r43-r74
13. Grantham, R. (1980) *Trends Biochem. Sci.* 5, 327-331
14. Trifonov, E.N. and Sussman, J.L. (1980) *Proc. Nat. Acad. Sci. USA* 77, 3816-3820
15. Eigen, M. and Winkler-Oswatitsch, R. (1981) *Naturwissenschaften* 68, 282-292
16. Sutcliffe, J.G., Shinnick, T.M., Green, N., Liu, F.-T., Niman, H.L., and Lerner, R.A. (1980) *Nature* 287, 801-805
17. Dijkema, R., Dekker, B.M.M. and Van Ormondt, H. (1980) *Gene* 9, 141-156
18. Waring, R.B., Davies, R.W., Lee, S., Grisi, E., Berks, M.M., and Scazzocchio, C. (1981) *Cell* 27, 4-11
19. Ohtsubo, H. and Ohtsubo, E. (1978) *Proc. Nat. Acad. Sci. USA* 75, 615-619
20. Ohtsubo, H., Nyman, K., Doroszkiewicz, W. and Ohtsubo, E. (1981) *Nature* 292, 640-643
21. Sugimoto, K., Oka, A., Sugisaki, H., Takanami, M., Nishimura, A., Yasuda, Y. and Hirota, Y. (1979) *Proc. Nat. Acad. Sci. USA* 76, 575-579
22. Meijer, M., Beck, E., Hansen, F.G., Bergmans, H.E.N., Messer, W., von Meyenburg, K. and Schaller, H. (1979) *Proc. Nat. Acad. Sci. USA* 76, 580-584
23. Lothar, H. and Messer, W. (1981) *Nature* 294, 376-378
24. Nakamura, M., Yamada, M., Hirota, Y., Sugimoto, K., Oka, A. and Takanami, M. (1981) *Nucleic Acids Res.* 9, 4669-4676
25. Brosius, J., Dull, T.J., Sleeter, D.D. and Noller, H.F. (1981) *J. Mol. Biol.* 148, 107-127
26. Csordas-Toth, E., Boros, I. and Venetianer, P. (1979) *Nucleic Acids Res.* 7, 2189-2197
27. Brosius, J., Palmer, M.L., Kennedy, P.J. and Noller, H.F. (1978) *Proc. Nat. Acad. Sci. USA* 75, 4801-4805
28. Brosius, J., Dull, T.J. and Noller, H.F. (1980) *Proc. Nat. Acad. Sci. USA* 77, 201-204
29. Spritz, R.A., Deriel, J.K., Forget, B.G. and Weissman, S.M. (1980) *Cell* 21, 639-646
30. Rubtsov, P.M., Musakhanov, M.M., Zakharyev, V.M., Krayev, A.S., Skryabin, K.G. and Bayev, A.A. (1980) *Nucleic Acids Res.* 8, 5779-5794
31. Hartley, J.L. and Donelson, J.E. (1980) *Nature* 286, 860-864
32. Hindley, J. and Phear, G.A. (1979) *Nucleic Acids Res.* 7, 361-375
33. Singleton, C.K., Roeder, W.D., Bogosian, G., Somerville, R.L. and Weith, H.L. (1980) *Nucleic Acids Res.* 8, 1551-1560



My NCBI  
[Sign In] [Regis]

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Book

Search PubMed for [ ] Go Clear

Limits Preview/Index History Clipboard Details

Display AbstractPlus Show 20 Sort by Send to

All: 1 Review: 0

☐ 1: Proc Int Conf Intell Syst Mol Biol. 1994;2:354-62.

Links

**The prediction of human exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames.**

**Solovyev VV, Salamov AA, Lawrence CB.**

Department of Cell Biology, Baylor College of Medicine, Houston, TX 77030, USA.

Discriminant analysis is applied to the problem of recognition 5'-, internal and 3'-exons in human DNA sequences. Specific recognition functions were developed for revealing exons of particular types. The method based on a splice site prediction algorithm that uses the linear Fisher discriminant to combine the information about significant triplet frequencies of various functional parts of splice site regions and preferences of oligonucleotides in protein coding and intron regions (Solovyev, Lawrence, 1994). The accuracy of our splice site recognition function is about 97%. A discriminant function for 5'-exon prediction includes hexanucleotide composition of upstream region, triplet composition around the ATG codon, ORF coding potential, donor splice site potential and composition of downstream intron region. For internal exon prediction, we combine in a discriminant function the characteristics describing the 5'-intron region, donor splice site, coding region, acceptor splice site and 3'-intron region for each open reading frame flanked by GT and AG base pairs. The accuracy of precise internal exon recognition on a test set of 451 exon and 246693 pseudoexon sequences is 77% with a specificity of 79% and a level of pseudoexon ORF prediction of 99.96%. The recognition quality computed at the level of individual nucleotides is 89% for exon sequences and 98% for intron sequences. A discriminant function for 3'-exon prediction includes octanucleotide composition of upstream intron region, triplet composition around the stop codon, ORF coding potential, acceptor splice site potential and hexanucleotide composition of downstream region. (ABSTRACT TRUNCATED AT 250 WORDS)

PMID: 7584412 [PubMed - indexed for MEDLINE]

Display AbstractPlus Show 20 Sort by Send to

[Write to the Help Desk](#)

### Related Links

Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames [Nucleic Acids Res. 1994]

Identification of human gene functional regions based on oligonucleotide composition [Proc Int Conf Intell Syst Mol Biol. 1993]

Identification of human gene structure using linear discriminant functions and dynamic programming [Proc Int Conf Intell Syst Mol Biol. 1995]

The prediction of exons through an analysis of spliceable open reading frames. [Nucleic Acids Res. 1992]

[Analysis, identification and correction of some errors of model refseqs appeared in NCBI Human Gene Database by in silico cloning and experimental verification of novel human genes] [Genes] [Xue Bao. 2004]

See all Related Articles...

[NCBI](#) | [NLM](#) | [NIH](#)  
Department of Health & Human Services  
[Privacy Statement](#) | [Freedom of Information Act](#) | [Disclaimer](#)

Aug 14 2006 08:07:58



A service of the National Library of Medicine  
and the National Institutes of Health

My NCBI

[\[Sign In\]](#) [\[Regis\]](#)

All Databases

PubMed

Nucleotide

Protein

Genome

Structure

OMIM

PMC

Journals

Book

Search  for

[Limits](#) [Preview/Index](#) [History](#) [Clipboard](#) [Details](#)

Display  Show  Sort by  Send to

All: 1 Review: 0

☐ 1: [J Biochem \(Tokyo\)](#). 1986 Jun;99(6):1579-90.

[Links](#)

### Determination of the initiation sites of transcription and translation of the *uvrD* gene of *Escherichia coli*.

**Yamamoto Y, Ogawa T, Shinagawa H, Nakayama T, Matsuo H, Ogawa H.**

Prior to the analysis of transcription and translation, the nucleotide sequence of the *uvrD* gene and its neighboring regions was determined by the method of Maxam and Gilbert (Maxam & Gilbert (1980) *Methods Enzymol.* 65, 499-560). Disagreement in 14 positions between the nucleotide sequence determined by us and that reported previously (Finch & Emmerson (1984) *Nucl. Acids Res.* 11, 5789-5799) was found. We reexamined these disputed regions. The initiation site of transcription of the *uvrD* gene was determined by analyzing the transcripts synthesized in vitro. It was found that transcription of the *uvrD* gene starts from the A nucleotide, which is the first one of the SOS box of the *uvrD*. The amino terminal sequence and the amino acid composition of the purified UvrD protein (helicase II) were determined. It was found that translation starts from the first ATG codon, which lies 77 nucleotides downstream from the initiation site of transcription. The amino acid composition of the purified UvrD protein agreed well with that deduced from the nucleotide sequence.

PMID: 2943729 [PubMed - indexed for MEDLINE]

Display  Show  Sort by  Send to

[Write to the Help Desk](#)

[NCBI](#) | [NLM](#) | [NIH](#)

[Department of Health & Human Services](#)

[Privacy Statement](#) | [Freedom of Information Act](#) | [Disclaimer](#)

### Related Links

Transcription of the *uvrD* gene of *Escherichia coli* is controlled by the *lexA* repressor [\[Mol Gen Genet. 1983\]](#)

The *E. coli uvrD* gene product is DNA helicase II. [\[Mol Gen Genet. 1983\]](#)

The molecular cloning of the gene encoding the *Escherichia coli* 75-kDa helicase and the determination of its nucleotide sequence and genetic map position. [\[J Biol Chem. 1989\]](#)

Structure of the *Bacillus sphaericus* R modification methylase [\[Proc Natl Acad Sci USA. 1983\]](#)

Nucleotide sequence of the *phoS* gene, the structural gene for the phosphate-binding protein of *Escherichia coli*. [\[J Bacteriol. 1984\]](#)

[See all Related Articles...](#)

Aug 14 2006 08:07:58



Login: Register

[Home](#) [Browse](#) [Search](#) [Abstract Databases](#) [My Settings](#) [Alerts](#) [Help](#)

**Quick Search** Title, abstract, keywords  Author  e.  
 search tips Journal/book title  Volume  Issue  Page

**Journal of Molecular Biology**

Volume 226, Issue 1, 5 July 1992, Pages 141-157

doi:10.1016/0022-2836(92)90130-C Cite or Link Using DOI  
Copyright © 1992 Published by Elsevier Ltd.

**This Document****► Abstract****Actions**

- Cited By
- Save as Citation Alert
- E-mail Article
- Export Citation
- Add to my Quick Links

**Article****Prediction of gene structure\*1**

Roderic Guigó<sup>†</sup>, Steen Knudsen<sup>§</sup>, Neil Drake<sup>||</sup> and Temple Smith<sup>†</sup>

Molecular Biology Computer Research Resource Dana-Farber Cancer Institute and Harvard School of Public Health 44 Binney St., Boston, MA 02215, U.S.A.

Received 4 September 1991; accepted 14 January 1992. Available online 2 November 2004.

**Abstract**

We have developed a hierarchical rule base system for identifying genes in DNA sequences. Atomic sites (such as initiation codons, stop codons, acceptor sites and donor sites) are identified by a number of different methods and evaluated by a set of filters and rules chosen to maximize sensitivity; these are combined into higher-order gene elements (such as exons), evaluated, filtered and combined as equivalence classes into probable genes, which are evaluated and ranked. The system has been tested on an extensive collection of vertebrate genes smaller than 15,000 bases. Results obtained show that, on average, 88% of the predicted coding region for a transcription unit is actually coding, and 80% of the actual coding is correctly predicted. This will, in most applications, be sufficient for a search against protein sequence databases for the identification of probable gene function. In addition, the system provides a general test platform for both gene atomic site identification and the rules for their evaluation and assembly.

**Author Keywords:** gene identification; exon structure; intron splicing; coding sequence; artificial intelligence

Corresponding author. Author to whom correspondence should be addressed.

\*1 This work was supported by National Library of Medicine grant LM05205 and by a postdoctoral fellowship from the Ministerio de Educación y Ciencia (Spain) to R.G. A beta version of the GeneId system has been made freely available to the research community by an automatic mail server.

† Present address: Bio-Molecular Engineering Research Center, Boston University, 36 Cummington St., Boston, MA 02215, U.S.A.

§ Present address: CEDB, University of West Florida, 11000 University Parkway, Pensacola FL 32514-5751, U.S.A.


|| Present address: Tufts University Medical Center, Boston, MA 03114, U.S.A.

---

**Journal of Molecular Biology**

Volume 226, Issue 1 , 5 July 1992, Pages 141-157

**This Document****► Abstract****Actions**

- Cited By
- Save as Citation Alert
- E-mail Article
- Export Citation
-  Add to my Quick Links

---

[Home](#) [Browse](#) [Search](#) [Abstract Databases](#) [My Settings](#) [Alerts](#) [Help](#)



[About ScienceDirect](#) | [Contact Us](#) | [Terms & Conditions](#) | [Privacy Policy](#)

Copyright © 2006 Elsevier B.V. All rights reserved. ScienceDirect® is a registered trademark of Elsevier B.V.



# Nucleic Acids Research

Journal List > Nucleic Acids Res > v.22(24); Dec 11, 1994

■ Summary

Selected References

Page Browse

PDF (1.6M)

Contents

Archive

Nucleic Acids Res. 1994 December 11; 22(24): 5156-5163.

[Copyright notice](#)

## Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames

V V Solovyev, A A Salamov, and C B Lawrence

Department of Cell Biology, Baylor College of Medicine, Houston, TX 77030.

Related material:

PubMed related arts

GO

► This article has been [cited by](#) other articles in PMC.

PubMed articles by:

Solovyev, V.

Salamov, A.

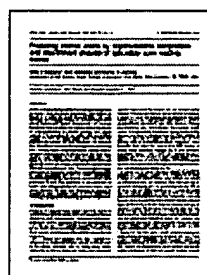
Lawrence, C.

### Abstract

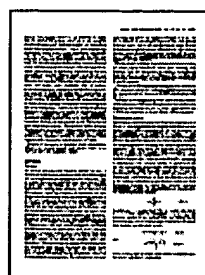
A new method which predicts internal exon sequences in human DNA has been developed. This method is based on a splice site prediction algorithm that uses the linear discriminant function to combine information about significant triplet frequencies of various functional regions and preferences of oligonucleotides in protein coding and intron regions. The splice site recognition function is 97% for donor splice sites and 96% for acceptor splice sites. For exon prediction, we combine in a discriminant function the characteristics of the coding region, donor splice site, coding region, acceptor splice site and 3'-intron region. The accuracy of precise internal exon prediction on a test set of 451 exon and 246693 pseudoexon sequences is 77% with a splice site recognition quality computed at the level of individual nucleotides is 89% for exons and 98% for intron sequences. This corresponds to a correlation coefficient for exon prediction of 0.87. The precision of this approach is better than other methods and has been tested on a large set of human cDNA sequences. We have also developed a means for predicting exon-exon junctions in cDNA sequences. This method can be useful for selecting optimal PCR primers.

### Full text

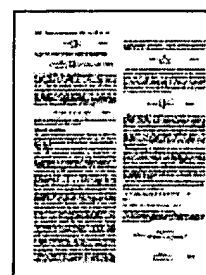
Full text is available as a scanned copy of the original print version. Get a preview of the [complete article](#) (1.6M), or see the PubMed citation or the full text of the article. Click on a page below to browse page by page.



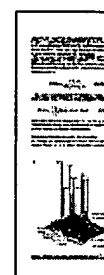
5156



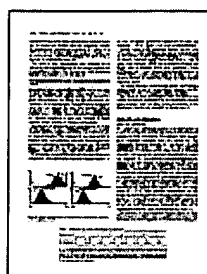
5157



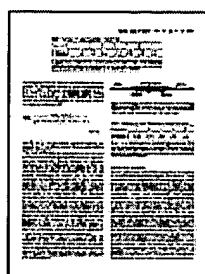
5158



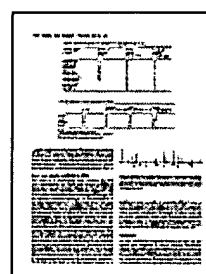
51



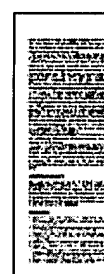
5160



5161



5162



51

## Images in this article



**Image**  
**on p.5159**

Click on the image to see a larger version.

## Selected References

This list contains those references that cite another article in PMC or have a citation in PubMed. I original references for this article.

- Guigó R, Knudsen S, Drake N, Smith T. Prediction of gene structure. *J Mol Biol.* 1992 Jun; [PubMed]
- Hutchinson GB, Hayden MR. The prediction of exons through an analysis of spliceable o *Nucleic Acids Res.* 1992 Jul 11;20(13):3453–3462. [PubMed]
- Snyder EE, Stormo GD. Identification of coding regions in genomic DNA sequences: an programming and neural networks. *Nucleic Acids Res.* 1993 Feb 11;21(3):607–613. [PubMed]
- Cinkosky MJ, Fickett JW, Gilna P, Burks C. Electronic data publishing and GenBank. *Science.* 1991 May 31;252(5010):1273–1277. [PubMed]



- Penotti FE. Human pre-mRNA splicing signals. *J Theor Biol.* 1991 Jun 7;150(3):385–420. [PubMed]
- Senapathy P, Shapiro MB, Harris NL. Splice junctions, branch point sites, and exons: sequence identification, and applications to genome project. *Methods Enzymol.* 1990;183:252–278.
- Stephens RM, Schneider TD. Features of spliceosome evolution and function inferred from information at human splice sites. *J Mol Biol.* 1992 Dec 20;228(4):1124–1136. [PubMed]
- Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage. *Biochim Biophys Acta.* 1975 Oct 20;405(2):442–451. [PubMed]
- Lawrence CB, Solovyev VV. Assignment of position-specific error probability to primary. *Nucleic Acids Res.* 1994 Apr 11;22(7):1272–1280. [PubMed]
- Staden R. Finding protein coding regions in genomic sequences. *Methods Enzymol.* 1990;183:15–32.
- Fickett JW, Tung CS. Assessment of protein coding measures. *Nucleic Acids Res.* 1992 I [PubMed]
- Uberbacher EC, Mural RJ. Locating protein-coding regions in human DNA sequences by network approach. *Proc Natl Acad Sci U S A.* 1991 Dec 15;88(24):11261–11265. [PubMed]
- Farber R, Lapedes A, Sirotkin K. Determination of eukaryotic protein coding regions using information theory. *J Mol Biol.* 1992 Jul 20;226(2):471–479. [PubMed]
- Nakata K, Kanehisa M, DeLisi C. Prediction of splice junctions in mRNA sequences. *Nucleic Acids Res.* 1985 Jul 25;13(14):5327–5340. [PubMed]
- Brunak S, Engelbrecht J, Knudsen S. Prediction of human mRNA donor and acceptor sites. *J Mol Biol.* 1991 Jul 5;220(1):49–65. [PubMed]
- Fields CA, Soderlund CA. gm: a practical tool for automating DNA sequence analysis. *Comput Appl Biosci.* 1990 Jul;6(3):263–270. [PubMed]

---

[Write to PMC](#) | [PMC Home](#) | [PubMed](#)  
[NCBI](#) | [U.S. National Library of Medicine](#)  
[NIH](#) | [Department of Health and Human Services](#)  
[Privacy Policy](#) | [Disclaimer](#) | [Freedom of Information Act](#)



A service of the National Library of Medicine  
and the National Institutes of Health

My NCBI

[Sign In] [Register]

All Databases

PubMed

Nucleotide

Protein

Genome

Structure

OMIM

PMC

Journals

Book

Search  for

Limits

Preview/Index

History

Clipboard

Details

Display  Show  Sort by  Send to

About Entrez

Text Version

Entrez PubMed

Overview

Help | FAQ

Tutorials

New/Noteworthy E-Utilities

PubMed Services

Journals Database

MeSH Database

Single Citation Matcher

Batch Citation Matcher

Clinical Queries

Special Queries

LinkOut

My NCBI

Related Resources

Order Documents

NLM Mobile

NLM Catalog

NLM Gateway

TOXNET

Consumer Health

Clinical Alerts

ClinicalTrials.gov

PubMed Central

☐ 1: [Nucleic Acids Res.](#) 1992 Jul 11;20(13):3453-62.

**FREE full text article**  
in PubMed Central

Lir

### The prediction of exons through an analysis of spliceable open reading frames.

**Hutchinson GB, Hayden MR.**

Department of Medical Genetics, University of British Columbia, Vancouver, Canada.

We have developed a computer program which predicts internal exons from naive genomic sequence data and which will run on any IBM-compatible 80286 (or higher) computer. The algorithm searches a sequence for 'spliceable open reading frames' (SORFs), which are open reading frames bracketed by suitable splice-recognition sequences, and then analyzes the region for codon usage. Potential exons are stratified according to the reliability of their prediction, from confidence levels 1 to 5. The program is designed to predict internal exons of length greater than 60 nucleotides. In an analysis of 116 genes of a training set, 384 out of 441 such exons (87.1%) are identified, with 280 (63.5%) of predictions matching the true exon exactly (at both 5' and 3' splice junctions and in the correct reading frame), and with 104 (23.6%) exons matching partially. In a similar analysis of 14 genes in a test set unrelated to the genes used to generate the parameters of the program, 70 out of 80 internal exons greater than 60 bp in length are identified (87.5%), with 47 completely and 23 partially matched. SORFs that partially match true internal exons share at least one splice junction with the exon, or share both splice junctions but are interpreted in an incorrect reading frame. Specificity (the percentage of SORFs that correspond to true exons) varies from 91% at confidence level 1 to 16% at confidence level 5, with an overall specificity of 35-40%. The output displays nucleotide position, confidence level, reading frame phase at the 5' and 3' ends, acceptor and donor sequences and scoring statistics and also gives an amino acid translation

### Related Links

Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames [Nucleic Acids Res. 1994

The prediction of human exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames [Intel Syst Mol Biol. 1994

[Analysis, identification and correction of some errors of model refseqs appeared in NCBI Human Gene Database by in silico cloning and experimental verification of novel human genes] [Mol Biotechnol. 2004





Recognizing exons in genomic sequence using GRAIL II [Genet Eng (N Y). 1994

Pombe: a gene-finding and exon-intron structure prediction system for fission yeast. [Yeast. 1998

See all Related Articles...

of the potential exon. SORFIND compares favourably with other programs currently used to predict protein-coding regions.

PMID: 1321415 [PubMed - indexed for MEDLINE]

Display   Show   Sort by   Send to  

[Write to the Help Desk](#)

[NCBI](#) | [NLM](#) | [NIH](#)

[Department of Health & Human Services](#)

[Privacy Statement](#) | [Freedom of Information Act](#) | [Disclaimer](#)

Aug 14 2006 08:07:58

[First Hit](#) [Fwd Refs](#)[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

Generate Collection

Print

L14: Entry 256 of 268

File: USPT

Dec 15, 1992

DOCUMENT-IDENTIFIER: US 5171844 A

**\*\* See image for Certificate of Correction \*\***

TITLE: Proteins with factor VIII activity: process for their preparation using genetically-engineered cells and pharmaceutical compositions containing them

Drawing Description Text (5):

a. The landmarks of the pSV2-derived vector: two tandemly situated promoters: the SV40 early transcription promoter (SVep) and the Rous Sarcoma Virus-Long Terminal Repeat (RSV-LTR); the capping site (cap site) and 5' end of the messenger RNA (mRNA); the cDNA insert bearing the full-length Factor VIII coding region with the start codon (ATG), the open reading frame and the stop codon (TGA); the 3' noncoding region of the mRNA with a short intron and the polyadenylation signal (polyA) derived from SV40 DNA (compare to FIG. 2).

[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

[First Hit](#) [Fwd Refs](#)[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

Generate Collection

Print

L14: Entry 255 of 268

File: USPT

Mar 2, 1993

DOCUMENT-IDENTIFIER: US 5190756 A

TITLE: Methods and materials for expression of human plasminogen variant

Detailed Description Text (74):

Mammalian expression vector pRK-tPA was prepared from pRK5 (described in EP 307,247, supra, where the pCIS2.8c28D starting plasmid is described in EP 278,776 published Aug. 17, 1988 based on U.S. Ser. Nos. 07/071,674 and 06/907,297) and from t-PA cDNA (Pennica et al., Nature, 301: 214 (1983)). The cDNA was prepared for insertion into pRK5 by cutting with restriction endonuclease HindIII (which cuts 49 pairs 5' of the ATG start codon) and restriction endonuclease BalI (which cuts 276 base pairs downstream of the TGA stop codon). This cDNA was ligated into pRK5 previously cut with HindIII and SmaI using standard ligation methodology (Maniatis et al., Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, New York, 1982). This construct was named pRK-t-PA, and is shown in FIG. 2.

[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

Plasminogen variants are defined as molecules in which the amino acid sequence of native Pg has been modified, typically by a predetermined mutation, wherein at least one modification renders the plasminogen resistant to proteolytic cleavage to its two-chain form. Amino acid sequence variants by Pg include, for example, deletions from, or insertions or substitutions of, residues within the amino acid Pg sequence shown in FIG. 1. Any combination of deletion, insertion and substitution may also be made to arrive at the final construct, provided that the final construct possesses the desired resistance to cleavage and biological activity. Obviously, it is preferred that the mutations made in the DNA encoding the variant Pg do not place the sequence out of reading frame and it is further preferred that they do not create complementary regions that could produce secondary mRNA structure (see, e.g., European Patent Publication No. 075,444).

. The human .alpha. subunit cDNA was engineered for expression by digesting the full-length clone with NcoI, which spans the start ATG, and HindIII, which cleaves in the 3' untranslated region 215 base pairs (bp) downstream of the TAA stop codon. A 5' SalI site and Kozak consensus sequence (27) was provided by synthetic oligonucleotides, and a 3' SalI site by attaching linkers as described above. The DNA sequence of the engineered .alpha. subunit cDNA clone, which is approximately 600 bp in length, is shown in Table 7. This was inserted into the XhoI site of the CLH3AXSV2DHFR expression vector (FIG. 2). The endogenous 5' untranslated region and 3' polyadenylation signal were removed from the cDNA clone in the process of engineering and therefore were supplied by vector sequences: the MT-I promoter and the simian virus 40 (SV40) early polyadenylation signal, respectively.

[First Hit](#) [Fwd Refs](#)[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

Generate Collection

Print

L14: Entry 248 of 268

File: USPT

May 10, 1994

DOCUMENT-IDENTIFIER: US 5310678 A

TITLE: Newcastle disease virus gene clones

Detailed Description Text (42):

Referring first to the F gene cDNA and proceeding in the 5' to 3' direction, the F.sub.o -coding region is though to extend from the proposed ATG start codon at nucleotides 47-49 to a TGA stop codon at 1706-1708. The cDNA encodes the F.sub.o polypeptide which is cleaved in vivo to F.sub.2, F.sub.1 (F.sub.2 being to the 5'-end of the F.sub.o gene cDNA, F.sub.1 to the 3'-end). Cleavage occurs at the C-terminal side of the arginine encoded by nucleotides 392-394. The amino acid sequence after the proposed cleavage site, viz that encoded by nucleotides 395-454, is the same as that of the 20 amino acids at the N-terminal of F.sub.1 determined by C. D. Richardson et al., supra. Beyond the end of the F.sub.1 -coding sequence is a non-coding portion corresponding to the 3' end of the mRNA which then terminates in a poly-A sequence at nucleotides 1787-1792.

[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)



Referring first to the F gene cDNA and proceeding in the 5' to 3' direction, the F.sub.o -coding region is though to extend from the proposed ATG start codon at nucleotides 47-49 to a TGA stop codon at 1706-1708. The cDNA encodes the F.sub.o polypeptide which is cleaved in vivo to F.sub.2, F.sub.1 (F.sub.2 being to the 5'-end of the F.sub.o gene cDNA, F.sub.1 to the 3'-end). Cleavage occurs at the C-terminal side of the arginine encoded by nucleotides 392-394. The amino acid sequence after the proposed cleavage site, viz that encoded by nucleotides 395-454, is the same as that of the 20 amino acids at the N-terminal of F.sub.1 determined by C. D. Richardson et al., supra. Beyond the end of the F.sub.1 -coding sequence is a non-coding portion corresponding to the 3' end of the mRNA which then terminates in a poly-A sequence at nucleotides 1787-1792.

43 The DNA sequence shows

The DNA sequence shows five significant potential asparagine-linked glycosylation sites in F.sub.o, one (NRT) in F.sub.2 at 299-307 and four (NKT, NTS, NIS and NNS) in F.sub.1 at 617-625, 1142-1150, 1385-1393 and 1457-1465. The NNT site near the C-end of F.sub.1 is considered insignificant since it lies in the region of the protein which does not cross the membrane.

- 44 The amino acid sequence of the HN polypeptide gene is shown with an ATG start codon at nucleotides 1915-1917 and a TAG stop codon at nucleotides 3646-3648; this is followed by a 177-nucleotide non-coding region which terminates in a poly-A sequence at the 3' end of the mRNA. The DNA sequence shows six potential glycosylation sites in HN, (NNS, NDT, NKT, NHT, NPT, NKT) at 2269-2277, 2935-2943, 3211-3219, 3355-3363, 3412-3420 and 3526-3534.
- 45 The non-coding region contains encodes a potential glycosylation site (NQT) at 3712-3720 and has a further TGA stop codon at 3757-3759, near the 3' end of the mRNA, which may provide an explanation for the origin of HN.sub.o in certain strains of NDV.
- 46 The HN proteins of the NDV strains Ulster and Queensland are known to be synthesised in a precursor form (HN.sub.o) which is cleaved to active HN by the removal of a C-terminal glycopeptide. These considerations suggest that the gene encoding the HN.sub.o precursor for the HN protein of certain avirulent NDV strains may differ from the genes encoding the HN proteins of more virulent strains of NDV by mutations generating a longer open reading frame and the consequent synthesis of a larger HN polypeptide.
- 47 Full length cDNA encoding the F and HN polypeptides

[First Hit](#)   [Fwd Refs](#)[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

Generate Collection

Print

L14: Entry 238 of 268

File: USPT

May 28, 1996

DOCUMENT-IDENTIFIER: US 5520911 A

TITLE: Variants of plasminogen activators and processes for their production

Detailed Description Text (160):

Plasmid pRK7 was used as the vector for generation of the t-PA mutants. This plasmid, described in EP 278,776 published Aug. 17, 1988, is identical to pRK5 (EP publication number 307,247 published 15 Mar. 1989), except that the order of the endonuclease restriction sites in the polylinker region between Cla I and Hind III is reversed. The t-PA cDNA (Pennica et al., Nature, 301: 214 (1983)) was prepared for insertion into the vector by cutting with restriction endonuclease Hind III (which cuts 49 base pairs 5' of the ATG start codon) and restriction endonuclease Bal I (which cuts 276 base pairs downstream of the TGA stop codon). This cDNA was ligated into pRK7 previously cut with Hind III and Sma I using standard ligation methodology (Maniatis et al., Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, New York, 1982). This construct was named pRK7-t-PA.

[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

[First Hit](#) [Fwd Refs](#)[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

Generate Collection

Print

L14: Entry 232 of 268

File: USPT

Mar 4, 1997

DOCUMENT-IDENTIFIER: US 5608036 A

TITLE: Enhanced secretion of polypeptides

Detailed Description Text (71):

BDNFopt3 was prepared with polymer-supported synthesis using standard phosphoramidite chemistry methods. Due to the length of BDNFopt3, the gene was synthesized as four separate segments: segment 1 is 104 bases and contains some 5' untranslated sequence corresponding to vector sequence, an XbaI restriction site, an ATG start codon, and the first 76 bases of the BDNFopt3 nucleic acid sequence; segment 2 contains the next 117 bases of BDNFopt3; segment 3 contains the next 107 bases of BDNFopt3; and segment 4 contains the remaining 57 bases of BDNFopt3 along with the TAA stop codon, a BamHI restriction site sequence, and 5 additional nucleotides. The segments were ligated together using standard ligation protocols. Prior to ligation, three oligonucleotides were hybridized to the BDNFopt3 gene fragments to ensure that the four gene segments would be ligated together in the proper order. Each of the three oligonucleotides used spans one of the BDNFopt3 gene fragment junctions. The nucleic acid sequence of each of these oligonucleotides is set forth below: ##STR1##

[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

[First Hit](#) [Fwd Refs](#)[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

End of Result Set



Generate Collection

Print

L15: Entry 26 of 26

File: USPT

Sep 29, 1992

DOCUMENT-IDENTIFIER: US 5151267 A

TITLE: Bovine herpesvirus type 1 polypeptides and vaccines

Detailed Description Text (104):

The gI gene maps between 0.422 and 0.443 genome equivalents (FIGS. 4 and 5), which is within the BHV-1 HindIII A fragment described by Mayfield et al. (1983), supra. A KpnI plus AccI partial digestion of the HindIII A fragment produces a 3255 base pair (bp) subfragment which contains the entire gI gene coding sequence. DNA sequence analyses placed an AccI site 20 bp 5' to the ATG start codon, while the KpnI site is 420 bp 3' to the TGA stop codon. This fragment was inserted into a synthetic DNA polylinker present between the EcoRI and SalI sites of PBR328 (i.e., ppo126, not shown) to produce pgB complete (FIG. 8). To this end, the AccI asymmetric end of the 3255 bp fragment was first blunted with Klenow enzyme and the gI fragment was then ligated to the HpaI plus KpnI sites of ppo126 to give pgB complete. HpaI and KpnI sites are within the polylinker of ppo126 and are flanked respectively by a BglII and a BamHI site. The gI gene was then transferred from pgB complete as a 3260 bp BglII+BamHI fragment to the BamHI site of the vaccinia virus insertion vector pGS20 (FIG. 9) to generate pgBvax (plasmid pGS20 with gI gene). Moss et al. in Gene Amplification and Analysis, Vol. 3, pp. 201-213 (Papas et al. eds. 1983).

[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

The gI gene maps between 0.422 and 0.443 genome equivalents (FIGS. 4 and 5), which is within the BHV-1 HindIII A fragment described by Mayfield et al. (1983), supra. A KpnI plus AccI partial digestion of the HindIII A fragment produces a 3255 base pair (bp) subfragment which contains the entire gI gene coding sequence. DNA sequence analyses placed an AccI site 20 bp 5' to the ATG start codon, while the KpnI site is 420 bp 3' to the TGA stop codon. This fragment was inserted into a synthetic DNA polylinker present between the EcoRI and SalI sites of PBR328 (i.e., ppol26, not shown) to produce pgB complete (FIG. 8). To this end, the AccI asymmetric end of the 3255 bp fragment was first blunted with Klenow enzyme and the gI fragment was then ligated to the HpaI plus KpnI sites of ppol26 to give pgB complete. HpaI and KpnI sites are within the polylinker of ppol26 and are flanked respectively by a BglII and a BamHI site. The gI gene was then transferred from pgB complete as a 3260 bp BglII+BamHI fragment to the BamHI site of the vaccinia virus insertion vector pGS20 (FIG. 9) to generate pgBvax (plasmid pGS20 with gI gene). Moss et al. in Gene Amplification and Analysis, Vol. 3, pp. 201-213 (Papas et al. eds. 1983).

[First Hit](#)[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

Generate Collection

Print

L26: Entry 1 of 2

File: PGPB

Jun 29, 2006

DOCUMENT-IDENTIFIER: US 20060142949 A1

TITLE: System, method, and computer program product for dynamic display, and analysis of biological sequence data

Description of Disclosure:

[0105] Some embodiments of biological sequence tools 212 may include another tool of pane 430 that may be available for analyzing a loaded or user selected sequence region for what is commonly referred to as an open reading or translation frame. Typically, for what are referred to as eukaryotes, three nucleotide bases typically code for each translated protein base. The three nucleotide bases are commonly referred to as a codon that may be read by a cell's translation machinery in what is commonly referred to as the translation or reading frame. Each sequence of DNA has six possible reading frames, three in each direction. Typically, only one reading frame codes for a protein and is referred to as the open reading frame. As is known to those of ordinary skill in the related art, the open reading frame typically begins with what is referred to as a start codon, and ends with a stop codon. The open reading frame analysis tool may be accessible by a user selection of ORF tab 1505 as illustrated in FIG. 15. Upon selection of tab 1505, ORF scale bar 1520 may be displayed in ORF selectable field 1510. In some implementations, the scale bar may represent a selectable minimum size of the ORF to be identified in loaded sequence 1407 or selected sequence such as, for instance, selected sequence 1430 of FIG. 14. User 101 may interactively select a value represented on scale bar 1520 by moving ORF scale tab 1525, via commonly used methods such as clicking and dragging with a mouse, to the desired position along scale bar 1520. In the illustrated implementation, scale bar 1520 may use a variety of different incremental scales, such as for instance numbers of base residues, as well as what is referred to by those of ordinary skill in the related art as kilobases, megabases, centimorgans, or other incremental value used for sequence measurement. In some embodiments, tab 1525 may be set to some default value that could correspond to an average ORF size or some other value. A selection of analyze ORF button 1515 instructs generator 210 to find one or more open reading frames in a loaded sequence or user selection of sequence, using the user selected criteria of scale tab 1525. GUI manager 211 may return the results to the user in a variety of formats that could include one or more colored boxes displayed in sequence coordinates pane 425 aligned with the one or more identified ORF's of sequence residues 1425.

[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

0049] In accordance with some implementations, some targets hybridize with probes and remain at the probe locations, while non-hybridized targets are washed away. These hybridized targets, with their tags or labels, are thus spatially associated with the probes. The term "hybridization" refers to the process in which two single-stranded polynucleotides bind non-covalently to form a stable double-stranded polynucleotide. The term "hybridization" may also refer to triple-stranded hybridization, which is theoretically possible. The resulting (usually) double-stranded polynucleotide is a "hybrid." The proportion of the population of polynucleotides that forms stable hybrids is referred to herein as the "degree of hybridization." Hybridization probes usually are nucleic acids (such as oligonucleotides) capable of binding in a base-specific manner to a complementary strand of nucleic acid. Such probes include peptide nucleic acids, as described in Nielsen et al, Science 254 1497-1500 (1991) or Nielsen Curr Opin Biotechnol, 10 71-75 (1999) (both of which are hereby incorporated herein by reference), and other nucleic acid analogs and nucleic acid mimetics. The hybridized probe and target may sometimes be referred to as a probe-target pair. Detection of these pairs can serve a variety of purposes, such as to determine whether a target nucleic acid has a nucleotide sequence identical to or different from a specific reference sequence. See, for example, U.S. Pat. No. 5,837,832, referred to and incorporated above. Other uses include gene expression monitoring and evaluation (see, e.g., U.S. Pat. No. 5,800,992 to Fodor, et al, U.S. Pat. No. 6,040,138 to Lockhart, et al, and International App No PCT/US98/15151, published as WO99/05323, to Balaban, et al), genotyping (U.S. Pat. No. 5,856,092 to Dale, et al), or other detection of nucleic acids. The '992, '138, and '092 patents, and publication WO99/05323, are incorporated by reference herein in their entireties for all purposes.

[0050] The present invention also contemplates signal detection of hybridization between probes and targets in certain preferred embodiments. See U.S. Pat. Nos. 5,143,854, 5,578,832, 5,631,734, 5,936,324, 5,981,956, 6,025,601 incorporated above and in U.S. Pat. Nos. 5,834,758, 6,141,096, 6,185,030, 6,201,639, 6,218,803, and 6,225,625, in U.S. Patent application 60/364,731 and in PCT Application PCT/US99/06097 (published as WO99/47964), each of which also is hereby incorporated by reference in its entirety for all purposes.

[0051] A system and method for efficiently synthesizing probe arrays using masks is described in U.S. patent application Ser. No. 09/824,931, filed Apr. 3, 2001, that is hereby incorporated by reference herein in its entirety for all purposes. A system and method for a rapid and flexible microarray manufacturing and online ordering system is described in U.S. Provisional Patent Application Ser. No. 60/265,103 filed Jan. 29, 2001, that also is hereby incorporated herein by reference in its entirety for all purposes. Systems and methods for optical photolithography without masks are described in U.S. Pat. No. 6,271,957 and in U.S. patent application Ser. No. 09/683,374 filed Dec. 19, 2001, both of which are hereby incorporated by reference herein in their entireties for all purposes.

[0052] As noted, various techniques exist for depositing probes on a substrate or support. For example, "spotted arrays" are commercially fabricated, typically on microscope slides. These arrays consist of liquid spots containing biological material of potentially varying compositions and concentrations. For instance, a spot in the array may include a few strands of short oligonucleotides in a water solution, or it may include a high concentration of long strands of complex proteins. The Affymetrix.RTM. 417.TM. Arrayer and 427.TM. Arrayer are devices that deposit densely packed arrays of biological materials on microscope slides in accordance with these techniques. Aspects of these and other spot arrayers are described in U.S. Pat. Nos. 6,040,193 and 6,136,269 and in PCT Application No PCT/US99/00730 (International Publication Number WO 99/36760) incorporated above and in U.S. patent application Ser. No. 09/683,298 hereby incorporated by reference in its entirety for all purposes. Other techniques for generating spotted arrays also exist. For example, U.S. Pat. No. 6,040,193 to Winkler, et al is directed to processes for dispensing drops to generate spotted arrays. The '193 patent, and U.S. Pat. No. 5,885,837 to Winkler, also describe the use of micro-channels or micro-grooves on a substrate, or on a block placed on a substrate, to synthesize arrays of biological materials. These patents further describe separating



reactive regions of a substrate from each other by inert regions and spotting on the reactive regions The '193 and '837 patents are hereby incorporated by reference in their entireties Another technique is based on ejecting jets of biological material to form a spotted array Other implementations of the jetting technique may use devices such as syringes or piezo electric pumps to propel the biological material It will be understood that the foregoing are non-limiting examples of techniques for synthesizing, depositing, or positioning biological material onto or within a substrate For example, although a planar array surface is preferred in some implementations of the foregoing, a probe array may be fabricated on a surface of virtually any shape or even a multiplicity of surfaces Arrays may comprise probes synthesized or deposited on beads, fibers such as fiber optics, glass, silicon, silica or any other appropriate substrate, see U.S. Pat. No. 5,800,992 referred to and incorporated above and U.S. Pat. Nos. 5,770,358, 5,789,162, 5,708,153 and 6,361,947 all of which are hereby incorporated in their entireties for all purposes Arrays may be packaged in such a manner as to allow for diagnostics or other manipulation in an all inclusive device, see for example, U.S. Pat. Nos. 5,856,174 and 5,922,591 hereby incorporated in their entireties by reference for all purposes

[0057] Methods and apparatus for signal detection and processing of intensity data are disclosed in, for example, U.S. Pat. Nos. 5,143,854, 5,578,832, 5,631,734, 5,800,992, 5,834,758, 5,856,092, 5,936,324, 5,981,956, 6,025,601, 6,090,555, 6,141,096, 6,185,030, 6,201,639, 6,207,960, 6,218,803, 6,225,625, in PCT Application PCT/US99/06097 (published as WO99/47964) incorporated above, and in U.S. Pat. Nos. 5,547,839, 5,902,723, 6,171,793, 6,207,960, 6,252,236, 6,335,824, 6,490,533, 6,472,671, 6,403,320, and 6,407,858 each of which is hereby incorporated by reference in its entirety for all purposes. Other scanners or scanning systems are described in U.S. patent application Ser. No. 09/682,837 filed Oct. 23, 2001, Ser. No. 09/683,216 filed Dec. 3, 2001, Ser. No. 09/683,217 filed Dec. 3, 2001, Ser. No. 09/683,219 filed Dec. 3, 2001, and Ser. No. 10/389,194, filed Mar. 14, 2003, each of which is hereby incorporated by reference in its entirety for all purposes.

[0058] The present invention may also make use of various computer program products and software for a variety of purposes, such as probe design, management of data, analysis, and instrument operation. See, U.S. Pat. Nos. 5,593,839, 5,795,716, 5,974,164, 6,090,555, 6,188,783 incorporated above and U.S. Pat. Nos. 5,733,729, 6,066,454, 6,185,561, 6,223,127, 6,229,911 and 6,308,170, hereby incorporated herein in their entireties for all purposes.

[0059] Scanner 185 provides data representing the intensities (and possibly other characteristics, such as color) of the detected emissions, as well as the locations on the substrate where the emissions were detected. The data typically are stored in a memory device, such as system memory 120 of user computer 100, in the form of a data file or other data storage form or format. One type of data file, such as image data file 212 shown in FIG. 2, typically includes intensity and location information corresponding to elemental sub-areas of the scanned substrate. The term "elemental" in this context means that the intensities, and/or other characteristics, of the emissions from this area each are represented by a single value. When displayed as an image for viewing or processing, elemental picture elements, or pixels, often represent this information. Thus, for example, a pixel may have a single value representing the intensity of the elemental sub-area of the substrate from which the emissions were scanned. The pixel may also have another value representing another characteristic, such as color. For instance, a scanned elemental sub-area in which high-intensity emissions were detected may be represented by a pixel having high luminance (hereafter, a "bright" pixel), and low-intensity emissions may be represented by a pixel of low luminance (a "dim" pixel). Alternatively, the chromatic value of a pixel may be made to represent the intensity, color, or other characteristic of the detected emissions. Thus, an area of high-intensity emission may be displayed as a red pixel and an area of low-intensity emission as a blue pixel. As another example, detected emissions of one wavelength at a particular sub-area of the substrate may be represented as a red pixel, and emissions of a second wavelength detected at another sub-area may be represented by an adjacent blue pixel. Many other display schemes are known. Two examples of image data are data files in the form \* dat or \* tif as generated respectively by Affymetrix.RTM. Microarray Suite or Affymetrix.RTM. GeneChip.RTM. Operating Software based on images scanned from GeneChip.RTM. arrays, and by Affymetrix.RTM. Jaguar.TM. software based on images scanned from spotted arrays.

[0060] Probe-Array Analysis Applications 199 Generally, a human being may inspect a printed or displayed image constructed from the data in an image file and may identify those cells that are bright or dim, or are otherwise identified by a pixel characteristic (such as color). However, it frequently is desirable to provide this information in an automated, quantifiable, and repeatable way that is compatible with various image processing and/or analysis techniques. For example, the information may be provided for processing by a computer application that associates the locations where hybridized targets were detected with known locations where probes of known identities were synthesized or deposited. Other methods include tagging individual synthesis or support substrates (such as beads) using chemical, biological, electro-magnetic transducers or transmitters, and other identifiers. Information such as the nucleotide or monomer sequence of target DNA or RNA may then be deduced. Techniques for making these deductions are described, for example, in U.S. Pat. No. 5,733,729 and in U.S. Pat. No. 5,837,832, noted and incorporated above.

addition to the above general procedures which can be used for preparing recombinant DNA molecules and transformed unicellular organisms in accordance with the practices of this invention, other known techniques and modifications thereof can be used in carrying out the practice of the invention. In particular, techniques relating to genetic engineering have recently undergone explosive growth and development. Many recent U.S. Pat. Nos. disclose plasmids, genetically engineering microorganisms, and methods of conducting genetic engineering which can be used in the practice of the present invention. For example, U.S. Pat. No. 4,273,875 discloses a plasmid and a process of isolating the same. U.S. Pat. No. 4,304,863 discloses a process for producing bacteria by genetic engineering in which a hybrid plasmid is constructed and used to transform a bacterial host. U.S. Pat. No. 4,419,450 discloses a plasmid useful as a cloning vehicle in recombinant DNA work. U.S. Pat. No. 4,362,867 discloses recombinant cDNA construction methods and hybrid nucleotides produced thereby which are useful in cloning processes. U.S. Pat. No. 4,403,036 discloses genetic reagents for generating plasmids containing multiple copies of DNA segments. U.S. Pat. No. 4,363,877 discloses recombinant DNA transfer vectors. U.S. Pat. No. 4,356,270 discloses a recombinant DNA cloning vehicle and is a particularly useful disclosure for those with limited experience in the area of genetic engineering since it defines many of the terms used in genetic engineering and the basic processes used therein. U.S. Pat. No. 4,336,336 discloses a fused gene and a method of making the same. U.S. Pat. No. 4,349,629 discloses plasmid vectors and the production and use thereof. U.S. Pat. No. 4,332,901 discloses a cloning vector useful in recombinant DNA. Although some of these patents are directed to the production of a particular gene product that is not within the scope of the present invention, the procedures described therein can easily be modified to the practice of the invention described in this specification by those skilled in the art of genetic engineering.

[First Hit](#) [Fwd Refs](#)[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

End of Result Set



Generate Collection

Print

L10: Entry 48 of 48

File: USPT

Mar 8, 1994

DOCUMENT-IDENTIFIER: US 5292658 A

TITLE: Cloning and expressions of Renilla luciferase

Detailed Description Text (63):

The pTZRLuc-1 crude supernatants were further characterized by SDS-PAGE. The Coomassie-stained gel contained numerous bands, one of which ran in the vicinity of native luciferase. To confirm that this band was recombinant lucifers, Western analysis was performed using rabbit polyclonal antibodies raised against native Renilla luciferase. The developed Western showed one band that migrated at the same position as native luciferase. No other products indicative of .beta.-galactosidase-luciferase fusion polypeptide were apparent, suggesting that either any putative fusion protein is in too low a concentration to be detected or, more likely, that no fusion protein is made. Though it has not been confirmed by DNA sequence analysis, any pTZRLuc-1 translation products initiating at the .beta.-galactosidase ATG start codon within the first three codons immediately adjacent to the first cDNA start codon may explain why we see IPTG induction of luciferase activity without production of a fusion product.

[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

[First Hit](#) [Fwd Refs](#)[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

Generate Collection

Print

L10: Entry 46 of 48

File: USPT

Jan 2, 1996

DOCUMENT-IDENTIFIER: US 5480972 A

TITLE: Allergenic proteins from Johnson grass pollen

Detailed Description Text (2):

The present invention provides nucleic acid sequences coding for Sor h I, a major allergen found in Johnson grass pollen. The nucleic acid sequence coding for Sor h I preferably has the sequence shown in FIG. 5 (SEQ ID NO: 1). Sequence analysis of the Sor h I clone 3S revealed that the cDNA insert is 1072 nucleotide long and contains 3 possible in-frame ATG start codons at nucleotide positions 25, 37 and 40. The ATG codon at position 40 is proposed as the site for translation initiation. This corresponds to an open reading frame of 783 nucleotides terminating with a TAA stop codon at position 823 and coding for a protein of 261 amino acids. See FIG. 5 (SEQ ID NO: 1 and 2). A host cell transformed with a vector containing the cDNA insert of clone 3S has been deposited with the American Type Culture Collection ATCC No. 69106 on Oct. 28, 1992.

[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

[First Hit](#) [Fwd Refs](#)[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

End of Result Set



Generate Collection

Print

L36: Entry 13 of 13

File: USPT

Sep 29, 1992

DOCUMENT-IDENTIFIER: US 5151267 A

TITLE: Bovine herpesvirus type 1 polypeptides and vaccines

Detailed Description Text (104):

The gI gene maps between 0.422 and 0.443 genome equivalents (FIGS. 4 and 5), which is within the BHV-1 HindIII A fragment described by Mayfield et al. (1983), supra. A KpnI plus AccI partial digestion of the HindIII A fragment produces a 3255 base pair (bp) subfragment which contains the entire gI gene coding sequence. DNA sequence analyses placed an AccI site 20 bp 5' to the ATG start codon, while the KpnI site is 420 bp 3' to the TGA stop codon. This fragment was inserted into a synthetic DNA polylinker present between the EcoRI and SalI sites of PBR328 (i.e., ppol26, not shown) to produce pgB complete (FIG. 8). To this end, the AccI asymmetric end of the 3255 bp fragment was first blunted with Klenow enzyme and the gI fragment was then ligated to the HpaI plus KpnI sites of ppol26 to give pgB complete. HpaI and KpnI sites are within the polylinker of ppol26 and are flanked respectively by a BglII and a BamHI site. The gI gene was then transferred from pgB complete as a 3260 bp BglII+BamHI fragment to the BamHI site of the vaccinia virus insertion vector pGS20 (FIG. 9) to generate pgBvax (plasmid pGS20 with gI gene). Moss et al. in Gene Amplification and Analysis, Vol. 3, pp. 201-213 (Papas et al. eds. 1983).

[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

[First Hit](#) [Fwd Refs](#)[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

Generate Collection

Print

L36: Entry 12 of 13

File: USPT

Aug 16, 1994

DOCUMENT-IDENTIFIER: US 5338683 A

TITLE: Vaccinia virus containing DNA sequences encoding herpesvirus glycoproteins

Detailed Description Text (45):

DNA sequence analysis revealed an open reading frame extending from nucleotide positions 300 to 3239 reading from left to right relative to the EHV-1 genome, i.e. the ATG start codon was contained in the BamHI-a/EcoRI fragment and the stop codon TAA was contained in the BamHI-i fragment (3,59).

[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)



[First Hit](#) [Fwd Refs](#)[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

End of Result Set



Generate Collection

Print

L36: Entry 13 of 13

File: USPT

Sep 29, 1992

DOCUMENT-IDENTIFIER: US 5151267 A

TITLE: Bovine herpesvirus type 1 polypeptides and vaccines

Detailed Description Text (104):

The gI gene maps between 0.422 and 0.443 genome equivalents (FIGS. 4 and 5), which is within the BHV-1 HindIII A fragment described by Mayfield et al. (1983), supra. A KpnI plus AccI partial digestion of the HindIII A fragment produces a 3255 base pair (bp) subfragment which contains the entire gI gene coding sequence. DNA sequence analyses placed an AccI site 20 bp 5' to the ATG start codon, while the KpnI site is 420 bp 3' to the TGA stop codon. This fragment was inserted into a synthetic DNA polylinker present between the EcoRI and SalI sites of PBR328 (i.e., ppol26, not shown) to produce pgB complete (FIG. 8). To this end, the AccI asymmetric end of the 3255 bp fragment was first blunted with Klenow enzyme and the gI fragment was then ligated to the HpaI plus KpnI sites of ppol26 to give pgB complete. HpaI and KpnI sites are within the polylinker of ppol26 and are flanked respectively by a BglII and a BamHI site. The gI gene was then transferred from pgB complete as a 3260 bp BglII+BamHI fragment to the BamHI site of the vaccinia virus insertion vector pGS20 (FIG. 9) to generate pgBvax (plasmid pGS20 with gI gene). Moss et al. in Gene Amplification and Analysis, Vol. 3, pp. 201-213 (Papas et al. eds. 1983).

[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

[First Hit](#)   [Fwd Refs](#)[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

Generate Collection

Print

L36: Entry 11 of 13

File: USPT

Jan 9, 1996

DOCUMENT-IDENTIFIER: US 5482713 A

TITLE: Equine herpesvirus recombinant poxvirus vaccine

Detailed Description Text (45):

DNA sequence analysis revealed an open reading frame extending from nucleotide positions 300 to 3239 reading from left to right relative to the EHV-1 genome, i.e. the ATG start codon was contained in the BamHI-a/EcoRI fragment and the stop codon TAA was contained in the BamHI-i fragment (3,59).

[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

[First Hit](#) [Fwd Refs](#)[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)

Generate Collection

Print

L36: Entry 10 of 13

File: USPT

Feb 13, 1996

DOCUMENT-IDENTIFIER: US 5491086 A

TITLE: Purified thermostable nucleic acid polymerase and DNA coding sequences from pyrodictium species

Detailed Description Text (77):

The DNA sequence analysis of pPab14 revealed an open reading frame of 803 amino acids having an ATG start codon at nucleotide position 869 and a TGA stop codon at nucleotide position 3280. The 5' end of the Pab gene was mutagenized with oligonucleotide primers AW397 (SEQ ID No. 5) and AW398 (SEQ ID No. 6) by PCR amplification (as described below). AW397 (SEQ ID No. 5) is forward primer which was designed to alter the Pab DNA sequence at the ATG start to introduce an NdeI restriction site. Primer AW397 (SEQ ID No. 5) also introduced mutations in the fifth and sixth codons of the Pab polymerase gene sequence to be more compatible with the codon usage of E. coli, without changing the amino acid sequence of the encoded protein. The reverse primer, AW398 (SEQ ID No. 6), was chosen to include a SpeI site corresponding to amino acid position 174. In addition, a KpnI site was introduced after the SpeI site.

[Previous Doc](#)[Next Doc](#)[Go to Doc#](#)